UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

A DEFENSE OF NONREDUCTIVE MENTAL CAUSATION

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

BY

ANDREW RUSSO
Norman, Oklahoma
2013

A DEFENSE OF NONREDUCTIVE MENTAL CAUSATION

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF PHILOSOPHY

BY

_____
Dr. Martin Montminy, Chair

_____
Dr. Reinaldo Elugardo

_____
Dr. Chris Swoyer

_____
Dr. Neal Judisch

_____
Dr. Lynn Devenport

For my parents
Dino and Christine

For Allison Hurst
Thank you for all the support and love

Table of Contents

List of Tables

List of Figures

Abstract

Mental causation is a problem and *not* just a problem for the nonphysicalist. One of the many lessons learned from Jaegwon Kim's writings in the philosophy of mind is that mental causation is a problem for the *nonreductive physicalist* as well. A central component of the common sense picture we have of ourselves as persons is that our beliefs and desires *causally explain* our actions. But the completeness of the "brain sciences" threatens this picture. If all of our actions are causally explained by neurophysiological events occurring in our brains, what causal role is left for our *reasons* and *motives*? It would seem that these brain events do all the causal work there is to do, thus robbing the mental of its efficacy altogether or else making it a merely superfluous or redundant causal factor. This essay presents a systematic treatment of this exclusion dilemma from the perspective of a nonreductive physicalist. I argue that both horns of this dilemma can be avoided if we ground mental causation in counterfactual dependence between distinct events and understand the mind-body relation as event realization. Although in the final analysis our actions are *overdetermined* by their mental and neurophysiological antecedents, this overdetermination is entirely unproblematic.

Introduction

Mental causation is a problem and *not* just a problem for the nonphysicalist. One of the many lessons learned from Jaegwon Kim's writings in the philosophy of mind is that mental causation is a problem for the *nonreductive physicalist* as well. A central component of the common sense picture we have of ourselves as persons is that our beliefs and desires *causally explain* our actions. But the completeness of the "brain sciences" threatens this picture. If all of our actions are causally explained by neurophysiological events occurring in our brains, what causal role is left for our *reasons* and *motives*? It would seem that these brain events do all the causal work there is to do, thus robbing the mental of its efficacy altogether or else making it a merely superfluous or redundant causal factor.

This essay presents a systematic treatment of the exclusion problem from the perspective of a *nonreductive physicalist*. Chapter 1 is a review of the underlying assumptions and arguments that generate exclusion worries. Specifically, I introduce the thesis of nonreductionism and some historically important considerations in its favor, such as arguments from multiple realizability and Saul Kripke's modal arguments. Additionally, I introduce the completeness and exclusion assumptions. Along the way, I present some distinctions that will be important throughout the remainder of this essay, such as the differences between weak and strong modal distinctness and physicalism and nonphysicalism. I end the chapter by presenting the exclusion problem as the dilemma of epiphenomenalism and causal overdetermination.

Chapter 2 discusses several of the more influential and interesting resolutions of the exclusion dilemma. The aim here is not to offer knock-down arguments against any of these views, but provide some reasons for why I think a better solution lies elsewhere. In particular, I approach the exclusion dilemma by making the following set of assumptions: (i) *strong realism* about mental predicates, (ii) mental causation and neurophysiological causation are *not* different kinds of causation, (iii) the assumption of completeness is plausibly true, (iv) mental and neurophysiological phenomena causally explain the same sorts of bodily phenomena, and (v) both type- and token-reductionism are false.

In Chapter 3, I deny that some of the traditional notions pervasive in the literature on the exclusion problem are germane to the real concern facing mental causation. In particular, I jettison talk of "causal sufficiency" and "causal relevance" and replace them with the idea of causation as a *difference-making* relation. I go on to discuss the positions of Stephen Yablo and Sydney Shoemaker, who argue that the exclusion problem can be solved given a particular understanding of causation. According to these authors, if causation is taken to be a *proportional relation*, then the threat of epiphenomenalism and overdetermination can be avoided. At the end of this chapter, I criticize this general approach and argue that, if consistently applied, the proportionality requirement would leave few of our pre-reflective causal judgments intact. The focus on causation as a difference-making relation is an improvement over causal sufficiency and relevance, but the failure of the proportionality constraint motivates a search for better approach.

In Chapter 4, I argue that causation as difference-making should be understood in terms of *counterfactual dependence*. Specifically, I argue that David Lewis's understanding of counterfactual dependence is a defensible and plausible sufficient condition for causation. Moreover, I argue for a position I call *counterfactualism*, which holds that counterfactual dependence between distinct events can vindicate mental causation. I defend this position against some recent criticisms offered by Jaegwon Kim, who argues that the counterfactualist cannot adequately distinguish between causal and pseudo-causal relations nor can they properly ground the mental causation which sustains agency. The underlying intuition motivating much of Kim's arguments against counterfactualism is that causation is a *productive or generative relation*. At the end of this chapter, I argue that a productive conception of causation is inconsistent with mental causation. I conclude that counterfactualism, or something very much like it, remains our only viable option for grounding the efficacy of our beliefs and desires.

In the final chapter, I discuss the problem of causal overdetermination. This horn of the exclusion dilemma is particularly troublesome for me as my conclusions in previous chapters entail that there is both a mental and neurophysiological cause of our bodily movements. For the most part, the assumption that overdetermination is bad and ought to be avoided is a dogma held by most philosophers. If overdetermination is bad, it must be bad *for a reason*. I explore three features of standard cases of overdetermination that have the potential to make overdetermination problematic for the nonreductive physicalist. These

problems can be avoided, however, if (a) we ground mental causation in counterfactual dependence between distinct events and (b) understand the mind-body relation as *event realization*. Although in the final analysis our actions are overdetermined by their mental and neurophysiological antecedents, this overdetermination is entirely unproblematic.

Chapter 1

The Philosophy of Mind and Mental Causation

Worries about the causal efficacy of the mind begin with the problem that forcefully presented itself to Descartes's dualistic picture of mind and body. In Section 1.1, I provide a brief characterization of this problem, noting that it is rooted in the Cartesian conception of minds as essentially non-extended substances. Although much contemporary thought on the problem has not been kind to Descartes, it has become evident that worries about mental causation have not disappeared with the rejection of his mind-body dualism. Section 1.2 begins the discussion of nonreductionism, the contemporary analog of Descartes's dualism. I cover some well-traveled ground by discussing two of the more influential arguments in support of nonreductionism, namely Hilary Putnam's argument from multiple realizability and Saul Kripke's distinctive brand of modal argument. This discussion is meant to serve two purposes. First, I want to make it clear how I understand these influential lines of thought. Second, I use this discussion to clarify the thesis of nonreductionism. In Section 1.3, I provide the formulation of nonreductionism that is best supported by Putnam's and Kripke's arguments and,

additionally, take the opportunity to draw distinctions between reductive physicalism, nonreductive physicalism, and nonphysicalism. The present essay pursues a resolution of the exclusion problem within the confines of a *nonreductive physicalism*.

The contemporary problem of mental causation is not a single problem. In Section 1.4, I present two well-known problems of mental causation that resemble Descartes's original problem in that each one is rooted in the nature of the mental. In short, the mental lacks causal efficacy because either it is anomalous or highly extrinsic. These problems respectively originate in Donald Davidson's important arguments for the anomalism of the mental and Hilary Putnam's and Tyler Burge's arguments for anti-individualism about mental content. Section 1.5 begins my discussion and presentation of the problem that will be the focus of this essay, the exclusion problem. Unlike the other contemporary problems of mental causation, the exclusion problem is born out of the thought that the physical rather than the mental is a certain kind of way. Specifically, the exclusion problem arises in the context of our contemporary scientific conception of the physical world as causally complete. Section 1.5.1 contains a discussion of this completeness assumption, making several important points clear. Section 1.5.2 presents and, to an extent, clarifies the standard formulation of the exclusion principle and distinguishes its role in generating worries about mental causation. In Section 1.5.3, I finally present the exclusion problem as a dilemma between mental epiphenomenalism or causal overdetermination.

Section 1.1: Descartes's Problem of Mental Causation

The problem of mental causation has transformed since Descartes's contemporaries wondered "how man's soul, being only a thinking substance, can determine animal spirits so as to cause voluntary action".[1] In the sixth *Meditation*, Descartes argues that minds and bodies are distinct substances, existents wholly independent of one another. His argument depends upon establishing that minds and bodies do not belong to each other's essence.[2] Very roughly, minds can do what minds do without bodies and bodies can do what bodies do without minds. For instance, a billiard ball with a specific velocity can cause another billiard ball to move in a particular direction with some determinate speed. One thing that bodies do is move and, so the argument goes, bodies can move without the help of minds.

In the second *Meditation*, Descartes purports to have shown that minds can have "purely intellectual perceptions" in which they cognize reality without the assistance of the senses. This is a critical moment; maybe the most critical in all of the *Meditations*. In addition to its consequences for the Aristotelian framework pervasive in most philosophical thought at the time, it plays a prominent role in establishing that minds are substances. For if minds can cognize reality without bodies, minds can do what minds do without the assistance of bodies. It is at this very point in the *Meditations*, the point at which he claims to have shown us how to

---

[1] *The Essential Descartes*, ed. M. Wilson (New York: New American Library, 1969): 373.

[2] See (Carriero 2009). My reading of Descartes's *Meditations* owes much to this wonderful book. However, only I should be blamed for mistakes in the specific interpretation of Descartes's argument that follows. Though I do intend it to be an accurate characterization of how he argues in the sixth *Meditation*, it is certain that my treatment here does not do justice to the subtleties of the text.

have "purely intellectual perceptions", that Descartes is assuming body and all of its characteristics are "chimeras". Therefore, it is possible that the mind does what minds do without any bodies existing at all. The essences of both minds and bodies are free from one another and so minds and bodies are distinct substances.

This distinctness of mental and physical substance is one aspect of the well-known radical dissimilarity of minds and bodies. Minds and bodies are substances each characterized by an essential property the other could not possibly have. Bodies, which are essentially extended and whose characteristics are specific ways of being extended, could not possibly have any mental properties. On the other hand, minds, which are essentially thinking and whose characteristics are specific ways of thinking, could not possibly have any bodily properties. This radical dissimilarity between minds and bodies is standardly understood to be the source of Descartes's problem of mental causation. Gassendi asks, "How could there be effort directed against anything, or motion set up in it, unless there is mutual contact between what moves and what is moved? And how can there be contact without a body…?".[3] The problem for Descartes appears to lie with the very nature of the mental, specifically that it is an essentially *non-extended* substance.

Section 1.2: The Thesis of Nonreductionism

Nowadays, most philosophers have given up on the idea that having a mind consists in having some nonphysical substance that underlies the possession of a

[3] *The Philosophical Writings of Descartes*, vol. 2, ed. J. Cottingham, R. Stoothoff, and D. Murdoch (Cambridge: Cambridge University Press, 1984): 236ff. Evident in Gassendi's remarks is a conception of causation that requires impact or contact between the relata. There is no wonder that Gassendi was puzzled as to how a Cartesian mind could direct effort on a body. See (Papineau 2001) for an interesting discussion of this point and its consequences.

unique family of properties. If substances are admitted into our contemporary ontology at all, only physical substances are permitted. Yet, dualism of a sort is still widely defended on philosophical grounds,

> (Nonreductionism) For every physical phenomenon *p* and mental phenomenon *m*, *p* is distinct from *m*.[4]

There are, however, two issues facing this formulation of Nonreductionism. First, it is not clear what it means for phenomena to be *distinct*. A recent essay by Stoljar (2008) notes that there are at least five different notions doing business under the heading of 'distinctness' (Stoljar 2008, 264). Second, it is not clear that the arguments advanced in support of Nonreductionism actually support it *as stated*. Instead, a similar though importantly different assumption has been the focus of the debate between reductionists and nonreductionists, one which claims that mental phenomena are distinct from neurophysiological phenomena. The assumption of Nonreductionism, then, both needs *disambiguated* and *properly restricted*.[5]

The debate between reductionists and nonreductionists concerns whether mental phenomena are *identical* with or *distinct* from physical phenomena. I shall consider it an uncontroversial fact about identity that x and y are identical only if there is a *symmetrical necessitation* relation that holds between x and y. This fact about the relation of identity is captured by (I):

---

[4] I am following (Yablo 1992b) in the terminology here. By 'phenomenon' I lump together both properties and events. I shall name properties with upper case letters (e.g., M, N, etc.), events with lower case letters (e.g., m, n, etc.), and phenomena with italicized lower case letters (e.g., *m*, *n*, etc.).

[5] Another and more tricky issue with Nonreductionism concerns the lumping together of properties and events. Are we justified in treating properties and events as the same or similar in the context of discussing mental causation? Perhaps but this ultimately depends on the correct picture of the relation between these two kinds of entities. Specifically, I think this issue concerns the nature and individuation of events, an issue I will try to remain neutral on as much as I can in this first chapter. See Ch. 3 Section 3.3 for more on this point.

(I) For any x and y, x = y only if (a) x necessitates y and (b) y necessitates x

The condition in (I) does not specify the metaphysical kinds of things it ranges over (e.g., properties, events, facts, states, dispositions, etc.) since it is meant to range over all of them. Furthermore, x *necessitates* y just in case: it is impossible for x to be present in the absence of y.[6] In the rest of this essay, I will be concerned exclusively with the distinctness of mental and physical phenomena so the relevant substitution instance is as follows: for any mental phenomenon *m* and physical phenomenon *p*, *m* = *p* only if (a) *m* necessitates *p* and (b) *p* necessitates *m*.

There are two important and influential lines of thought meant to show that mental phenomena are distinct from physical phenomena. The first line of thought originates with the arguments from multiple realizability given by Putnam (1975b). According to these arguments, mental phenomena are distinct from physical phenomena because the (a)-condition in (I) fails to be satisfied, that is, mental phenomena do not necessitate physical phenomena. According to Stoljar (2008), this amounts to mental and physical phenomena being weakly modally distinct where x is *weakly modally distinct* from y just in case it is possible that x is present and y absent or it is possible that y is present and x absent (but not both) (Stoljar

[6] A few more remarks are in order. First, the notions of "presence" and "absence" are intended to be neutral with respect to different ways of being. Objects *exist*, events *occur*, and properties are *possessed* or *instantiated*. Second, the modality here should be read as *metaphysical* and not, e.g., conceptual or nomological. Third, necessitation is *temporally invariant* in sense that it is stable across all the instants of time within a possible world. If x necessitates y, then it will do so for every possible time t in a world. Fourth, I assume necessitation is *non-causal,* since x and y are present simultaneous. Fifth, necessitation is a *dependence* relation. Assuming that numbers are necessary entities it follows that any contingent entity whatsoever necessitates the number 7, since it is the case that necessarily, if that contingent entity exists, 7 exists. But it appears strange to think of the number 7 as depending upon this contingent entity. So, the necessitation relations under consideration here are to be understood as *non-causal dependency* relations.

2008, 265). Therefore, Putnam's argument from multiple realizability should be understood as supporting the following version of Nonreductionism:

> (Weak Nonreductionism) For every physical phenomenon *p* and mental phenomenon *m*, *p* is weakly modally distinct from *m*

The second line of thought derives from the modal arguments developed by Kripke (1980). Unlike the arguments from multiple realizability, modal arguments purport to show that mental phenomena are distinct from physical phenomena because *both* the (a)- and (b)-conditions in (I) fail to be satisfied. Stoljar (2008) labels this strong modal distinctness where x is *strongly modally distinct* from y just in case it is possible that x is present and y absent and it is possible that y is present and x absent (Stoljar 2008, 266). Kripke's modal arguments, then, support a much stronger version of Nonreductionism:

> (Strong Nonreductionism) For every physical phenomenon *p* and mental phenomenon *m*, *p* is strongly modally distinct from *m*

In the next two sections, I discuss these two influential lines of thought more carefully. Not only do I want to keep in mind that they support two different versions of Nonreductionism, I also want to emphasize that neither support the assumption that mental phenomena are distinct from physical phenomena *tout court*. Instead, mental phenomena are taken to be distinct from the physical phenomena *cited in or subsumed by the laws of the lower-level sciences*. Specifically, the arguments are intended to show that mental phenomena are either weakly or strongly modally distinct from *neurophysiological phenomena*. This restriction to the assumption of Nonreductionism is absolutely crucial to properly

understand the exclusion problem, since neurophysiological phenomena are standardly taken to be the primary causal competitors of the mental phenomena we consider to be among the causes of human behavior.

Section 1.2.1: Multiple Realizability

A corollary of (I) is that property P is identical to property Q only if P necessitates Q. Putnam's (1975b) considerations from *multiple realizability* are best understood as establishing that mental properties do not necessitate the physical properties cited in or subsumed by the laws of neurophysiology. In short, mental properties are multiply realizable properties and, in virtue of this, do not necessitate neurophysiological properties. The multiple realizability of mental properties is inconsistent with the satisfaction of the (a)-condition in (I) and so guarantees their *weak* modal distinctness from *neurophysiological* properties.

Putnam's argument begins with the observation that satisfying the (a)-condition is in tension with empirical facts about how mental properties are instantiated *in the actual world*. A variety of creatures that are mentally quite similar are neurophysiological quite distinct. He claims it is reasonable to believe on the empirical evidence we presently possess that, for any mental property M, M can be instantiated in some nomologically possible creature without that creature instantiating neurophysiological property N. In other words, it is consistent with the laws of the actual world that a creature exist which instantiates M without N. This nomological possibility shows that the (a)-condition is not satisfied. If it is

nomologically possible to instantiate M without N, it is metaphysically possible to instantiate M without N and so M does not necessitate N.

In addition, Jerry Fodor's (1974) discussion introduces what John Bickle (1998) has called multiple realizability "in a token system over times". During a single creature's mental history, it is possible for it to possess mental property M and neurophysiological property N at t_1 and that same mental property without N at t_2. Again, if this is nomologically possible (and it seems to be given our present evidence), it is metaphysically possible to instantiate M without N and so M does not necessitate N.

The typical conclusion drawn from these discussions is that mental properties are not identical with neurophysiological properties, but instead are *realized* in them.[7] The reason is that mental properties do not necessitate neurophysiological properties, since it is nomologically possible for there to exist creatures that possess mental properties and either possess neurophysiological properties very different from ours (e.g., octopi, reptiles, etc.) or possess no neurophysiological properties at all (e.g., Martians, androids, highly advanced artificial intelligences, etc.). The following has become a widely accepted condition on the relationship between mental and neurophysiological properties:

> (MR) Necessarily, for every mental property M and every neurophysiological property N which *realizes* M, it is nomologically possible that something instantiates M but not N.

[7] Kim (1993a), Sober (1999), and Shapiro (2000) present some reasons to think that these are the wrong conclusions to draw.

Putnam's discussion of multiple realizability strongly suggests that the relationship between mental and neurophysiological properties must be consistent with the nomological possibility of creatures that are mentally exactly similar to us although neurophysiological quite distinct.

Additionally, the considerations from multiple realizability have done much to undermine the related project of *globally reducing* psychological theory to neurophysiological theory. [8] Specifically, the nomological possibility of instantiating M without P is inconsistent with the common philosophical account of inter-theoretic reduction. The account of inter-theoretic reduction introduced by Nagel (1961) is standardly interpreted to require nomic equivalences or *bridge laws* between mental and neurophysiological properties. It is these nomic equivalences that are inconsistent with mental-neurophysiological realization. Fodor (1974) gives us the classic statement of global theory reduction in which the importance of these equivalences is apparent. Presumably, theories are at least interrelated nomic generalizations, or laws, of the form

(1) $M_1x \rightarrow M_2x$

where (1) should be read as the instantiation of property M_1 is nomologically sufficient for the instantiation of property M_2. A theory M_T is reduced to a theory P_T just in case each nomic generalization of M_T is reduced to a nomic generalization of P_T. The reduction of a psychological theory to a neurophysiological theory proceeds from the premise that each nomic

[8] See (Lewis 1980) and (Kim 1993a) for local reductions of psychology to neurophysiological theory.

generalization of the psychological theory is reducible to a nomic generalization of the neurophysiological theory. Theory reduction, then, is premised upon law reduction. A law such as (1) is reduced to a law such as (2)

(2) $N_1x \rightarrow N_2x$

just in case the following pair of bridge laws hold

(3) $M_1x \leftrightarrow N_1x$

(4) $M_2x \leftrightarrow N_2x$

where (3) and (4) should be understood as stating nomic equivalences between properties. From the law (2) and the bridge laws (3) and (4), we can logically derive (1), a law of M_T. Therefore, a theory M_T is reducible to a theory P_T just in case each nomic generalization of M_T is logically derivable from the nomic generalizations of P_T *with the help of* bridge laws like (3) and (4). Considerations from multiple realizability prevent the existence of being bridge laws of the form of (3) and (4), which prevents a global reduction of psychological theory to neurophysiological theory.

Section 1.2.2: Modal Arguments

The other family of reasons in support of the distinctness of mental and neurophysiological phenomena consists of modal arguments.[9] Modal arguments are, in a few important ways, different from considerations of multiple realizability. First, modal arguments attempt to establish more than Putnam's arguments from

[9] Some other well-known considerations are explanatory gap arguments and knowledge arguments. Although I have not worked out the details fully, I think modal arguments have a kind of primacy when it comes to establishing the distinctness of phenomena. I suspect that both explanatory gap arguments and knowledge arguments depend on the success of modal arguments.

multiple realizability. These arguments purport to show that mental phenomena are strongly modally distinct from neurophysiological phenomena, since neither the (a)- and (b)-conditions in (I) is satisfied. Second, as the discussion that follows will make clear, modal arguments make explicit appeal to *conceivability evidence*. These arguments begin by appealing to the conceivability of some state of affairs, whereas considerations from multiple realizability proceed from the claim that some states of affairs are nomologically possible. To claim that P is conceivable is to make the epistemic claim that P is not *a priori* false. This is much weaker than the empirical premise from which claims of multiple realizability are based, namely that certain states of affairs are consistent with the laws of the actual world. Third, unlike multiple realizability arguments which have tended to focus exclusively on *properties*, modal arguments have been developed for both properties and events.[10]

Let's begin with the modal argument against mental-physical property identity as it is developed by Kripke (1980). Where both 'C-fiber firings' and 'pain' rigidly denote some property, suppose for *reductio* that 'C-fiber firings = pain' is true. Kripke (1971) argues that if this identity statement is true, it is necessarily true. However, we can conceive of situations in which C-fiber firing is separable from pain:[11]

Disembodiment: there is pain without C-fiber firing

[10] I see no barrier to multiple realizability arguments applying to events. See (Yablo 1992b) for such considerations. I point this out only to make salient the historical fact that talk about multiple realizability has focused upon properties and not on other kinds of things.

[11] Modal arguments begin by establishing what David Chalmers (2002) calls an "epistemic gap" between the mental and the physical. Establishing this epistemic gap is suppose to preclude the possibility of reductive explanations of the mental and some argue that all physicalists require such explanations (see for instance (Jackson 2007)).

Zombies: there is C-fiber firing without pain

According to Kripke, if there is no way to explain away the apparent separability of pain and C-fiber firing, it is metaphysically possible to have one without the other. First, if there is no way to explain away the conceivability of disembodiment, it is metaphysically possible to have pain without C-fiber firing and so the mental property pain does not necessitate this neurophysiological property. This is the same conclusion reached by considerations from multiple realizability although it is reached via *inter alia* a weaker epistemic premise. Second, if there is no way to explain away the conceivability of zombies, it is possible to have C-fiber firings without pain. In addition to the claim that mental properties do not necessitate neurophysiological properties, this kind of modal argument purports to show that this neurophysiological property does not necessitate the mental property pain. At best, if there is no way to explain away the apparent separability of pain and C-fiber firing, then 'C-fiber firings = pain' is contingent. But since 'C-fiber firings = pain' is necessarily true if it is true at all, the identity statement 'C-fiber firings = pain' must be false. The question now is whether there is a way to explain away these admittedly conceivable scenarios.

In most cases of identifying a commonsense natural kind with a theoretical kind (e.g., heat and molecular motion), we can explain away the apparent separability by telling the following kind of story. When we think we are conceiving of heat without molecular motion, what we are really conceiving of is something else entirely. What we are conceiving of is *the feeling of heat* (viz., the

sensation of heat) as separable from molecular motion. That is, we pick out the phenomenon of heat by one of its contingent properties – that it tends to cause certain kinds of sensations in creatures like us. Similarly, when we think we conceive of molecular motion without heat, what we are really conceiving of is molecular motion not bringing about the feeling of heat in us. In either case, we are not really conceiving of the separability of heat and molecular motion.

In the case of pain and C-fiber firing, this strategy is not applicable (though Kripke admits this is no proof that no such strategies are available).[12] When we conceive of disembodiment (i.e., pain without C-fiber firing), we are conceiving of the feeling of pain without C-fiber firing. But conceiving of the feeling of pain is just to conceive of pain, since we pick out pain by one of its essential properties, namely its "immediate phenomenological quality". Similarly, when we conceive of zombies (i.e., C-fiber firing without pain), we are conceiving of C-fiber firing without the feeling of pain and, again, the absence of the feeling of pain just is the absence of pain. Overall, since being in pain is a conscious mental condition, there can be no distinction between the feeling and the reality or, as Kripke puts it, "the notion of an epistemic situation qualitatively identical to one in which the observer had a sensation S simply is one in which the observer had that sensation" (Kripke 1980, 152). The conclusion is that since we cannot explain away the apparent separability of pain and C-fiber firing, they really are separable and hence neither necessitates the other. The identity statement 'C-fiber firings = pain', then, can be

[12] See (Hill 1997) for an alternative strategy.

at most a contingent truth. But it is necessarily true if it is true at all so it is false that pain is identical C-fiber firing.

Various kinds of modal arguments have also been advanced in support of the distinctness of mental and physical events.[13] Kripke's original argument begins by claiming that if mental event m is identical to neurophysiological event n, then, by Leibniz's Law, m and n must share all of their properties, including their *de re* modal properties. One *de re* modal property of neurophysiological event n is the property *being essentially a brain state*. Kripke continues,

> Indeed, even more is true: not only being a brain state, but even being a brain state of a specific type is essential to n. The configuration of brain cells whose presence at a given time constitutes the presence of n at that time is essential to n, and in its absence n would not have existed (minor re-lettering) (Kripke 1980, 147 – 148).

If the mental event m is identical to neurophysiological event n, it "could not have existed without a quite specific type of configuration of molecules" (Kripke 1980, 148). *Prima facie*, m could have occurred without the existence of this very specific configuration of molecules and so is not *essentially a brain state of that type*. Therefore, m and n differ in their *de re* modal properties and must be distinct.[14]

---

[13] If one gives a fine-grained individuation of events along the lines of Kim (1973, 1976) in which events have constitutive properties, the distinctness of the constitutive properties entails the distinctness of the events. This kind of individuation of events justifies a unified treatment of properties and events in the context of discussing mental causation. See fn. 5.

[14] Token-identity theorists need to resist this kind of modal reasoning about events. A common reply is as follows: what is clear is that *some* mental event could have occurred without the existence of this very specific configuration of molecules, but it is not so clear that it is *the very mental event* the token-identity theorist claims is identical to this neurophysiological event. Yablo (1992b) reminds us, "Among the lessons of *Naming and Necessity* is that to find a thing x capable of existing in some counterfactual condition, one imagines this *directly* – as opposed to imagining

A similar argument can be run in the reverse. Mental event m has the property *being essentially a mental state*. Indeed, even more is true: not only being essentially a mental state, but *being essentially a mental state of a specific type*. If m is an intentional mental event, then the type of propositional attitude it is (e.g., belief, desire, etc.) and its propositional content are essential to m or, if it is phenomenal, its phenomenological character is essential. If the neurophysiological event n is identical to mental event m, it could not have occurred without being a specific propositional attitude having a specific propositional content or having a specific phenomenological character. *Prima facie*, the neurophysiological event n could have occurred without being this attitude with this content or without having this character. Thus, n is not *essentially a mental state of that type*. Therefore, n and m differ in their *de re* modal properties and must be distinct.[15]

Tyler Burge (1979) advances an argument along these lines that appeals to content properties. Intentional mental events have their content essentially. Mental content should be analyzed in terms of relational properties the thinker has with respect to his or her physical and/or social environment (viz., some kind of anti-

---

something y in that condition whose transworld identity with x must then be established. This is crucial if imaginability is to be a source of knowledge about *de re* possibility" (Yablo 1992b, 269, fn. 50). Yablo goes on to argue that the question here boils down to whether m really is imaginable in the absence of n or whether what one imagines is a distinct but similar m in n's absence (Yablo 1992b, 269, fn. 50). Here I follow Kripke and Yablo and claim that we are imagining m and not a distinct but similar event *and* that this is *prima facie* plausible.

[15] Yablo (1992b) questions the second part of this argument. Roughly, *if mental event m is taken to supervene on neurophysiological event p*, the essential physical properties of p must necessitate the essential mental properties of m. It follows, then, that p has the essential mental properties of m essentially (Yablo 1992b, 268). The point, however, is not to argue for a token identity theory but note an asymmetry in the above arguments. According to Yablo, mental and neurophysiological events are distinct, since "the essences of mental events are physically impoverished" (Yablo 1992b, 268) though the reverse is not the case.

individualism is true). It follows that mental events bear certain kinds of relations to the outside world essentially. *Prima facie*, neurophysiological event n does not bear these same relations to the outside world essentially. Therefore, m and n have different essential properties and so cannot be identical. As Bennett (2003) notes, nothing in this argument hangs on the assumption that some kind of anti-individualism about content is true. Suppose that some kind of individualist functional or conceptual role semantics is the true story about mental content. Given that intentional mental events have their content essentially, it follows that mental events bear certain kinds of relations to other mental events essentially. *Prima facie*, neurophysiological event n does not bear these same relations to mental events essentially. Therefore, m and n have different essential properties and so cannot be identical.

These modal arguments conclude that there is a more robust distinctness between mental and neurophysiological phenomena than what follows from multiple realizability. The reason is that it is consistent with (MR) that neurophysiological phenomena necessitate mental phenomena, but modal arguments purport to undermine even this. The difference is that multiple realizability supports the *weak modal distinctness* of mental and neurophysiological phenomena, since *one* of the two conditions stated in (I) fails to be satisfied (i.e., the (a)-condition), while modal arguments support the *strong modal distinctness* of mental and neurophysiological phenomena, since *both* conditions stated in (I) fail to be satisfied.

Section 1.3: The Thesis of Nonreductionism, Again

If we make sure to separate weak from strong modal distinctness *and* recognize the restriction to neurophysiological phenomena, multiple realizability and modal arguments support two different assumptions respectively:

> (Weak Restricted Nonreductionism) For every neurophysiological phenomenon *n* and every mental phenomenon *m*, *n* is weakly modally distinct from *m*
>
> (Strong Restricted Nonreductionism) For every neurophysiological phenomenon *n* and every mental phenomenon *m*, *n* is strongly modally distinct from *m*

The history I have emphasized shows that the philosophical focus has not been on physical phenomena *tout court* as bad candidates for mental phenomena, but only those phenomena cited in or subsumed by laws of the lower-level sciences with specific emphasis on neurophysiological phenomena. This has much in the way of a philosophical precedent. Both multiple realizability and modal arguments were, at the time of their original formulations, leveled at philosophers sympathetic to an identity theory of the mind. Theorists like U.T. Place (1956), Herbert Feigl (1958), J.J.C. Smart (1959) and even David Lewis (1966, 1972) and David Armstrong (1981) argued that mental phenomena are not just some physical phenomena or other, but specific kinds of physical phenomena, namely the neurophysiological phenomena assumed to be cited in or subsumed by neurophysiological laws. The influence of these arguments against specific reductionist projects eventually led to one or the other restricted version of nonreductionism gaining favor with the philosophical majority.

Furthermore, much of the motivation for arguing against these reductionist projects is to preserve the autonomy of psychology as a science. If mental phenomena are neurophysiological phenomena, nomic equivalences between mental and neurophysiological properties will obtain and so bridge laws will be available for a reduction of psychology to neurophysiology. Given such a reduction, the generalizations captured by psychologists will be generalizations already captured by the neurophysiologists or, at least, possibly captured by them. Psychologists might as well just become neurophysiologists. But accepting one or the other restricted version of nonreductionism does nothing to undermine this motivation. As long as mental phenomena turn out to be weakly modally distinct[16] from the physical phenomena cited in or subsumed by some lower-level science, psychology can continue on with the assumption that its generalizations are not in principle subject to theoretical reduction. No one is in the business of finding psychological generalizations besides the psychologist. As it turns out, the autonomy of psychology is consistent with mental phenomena being physical – they just cannot be, e.g., *neurophysiological* phenomena.

This lesson about the assumption of Nonreductionism has been recently stressed by Bennett (2008) and I believe it is one worth remaining clear on. The kind of physicalism that will be the focus of this essay is the *nonreductive physicalism* prevalent throughout much of analytic philosophy. It would misconstrue nonreductive physicalism to consider it as committed to the claim that

[16] It should be rather straightforward that strong modal distinctness entails weak modal distinctness. If (P & Q) is true then P is true by simplification which is enough to make (P ∨ Q) is true.

mental phenomena are (weakly or strongly modally) distinct from physical phenomena *tout court*. Indeed, the nonreductive physicalist is only committed to denying that mental phenomena are physical in, what Bennett (2008) calls, the narrow sense. That is, the nonreductive physicalist denies that mental phenomena are identical to any *neurophysiological* phenomena. My discussion of nonreductionism will, therefore, continue to suppose that one or the other restricted version of Nonreductionism is the relevant assumption for generating the exclusion problem. But now the question remains: Which restricted version must the *nonreductive physicalist* accept?

Section 1.3.1: Physicalism and Global Supervenience

Answering this question is facilitated by looking at what makes the *physicalist* different from the *nonphysicalist*. Their disagreement hinges on the truth of the following supervenience thesis:

> (Global Supervenience) Any minimal physical duplicate of the actual world is a duplicate *simpliciter* of the actual world.

A *minimal* physical duplicate of the actual world is any world that has precisely the same distribution of physical properties and particulars as the actual world *and nothing more*. Alternatively, we can put the thesis as follows. If P reports all the physical facts and P* reports all the facts, then if physicalism is true, 'if P then P*' is a necessary truth.[17] If M reports a subset of the facts reported by P*, namely all

[17] See (Lewis 1983), (Chalmers 1996), and (Jackson 1998). In other words, if someone accepts that 'if P then P*' is a necessary truth, then they hold that the physical way the actual world is *metaphysically necessitates* all the facts about the actual world. This should not be understood to include the epistemic claim that this necessitation is knowable only *a posteriori* (see (Chalmers 1996, 69). Additionally, there should be no further presumption that the necessitation is knowable

of the intentional and phenomenal facts, then the necessary truth 'if P then P*' entails that 'if P then M' is also necessary truth. In other words, Global Supervenience implies that any minimal physical duplicate of the actual world is a psychological duplicate of the actual world.

On the other hand, the nonphysicalist denies this and maintains that there are minimal physical duplicates of the actual world that are not duplicates *simpliciter* of the actual world. Specifically, they uphold that it is possible for there to be psychological differences in a world that is a minimal physical duplicate of the actual world. In other words, if nonphysicalism is true, then 'if P then M' is merely a contingent truth. The following table represents these differences:

	Global Supervenience
Nonphysicalism	(-) There are minimal physical duplicates of @ that are not duplicates *simpliciter* of @
Physicalism	(+) Minimal physical duplicates of @ are duplicates *simpliciter* of @

Table 1: Physicalism and Nonphysicalism

Not only must the *nonreductive* physicalist differentiate themselves from the nonphysicalist by accepting Global Supervenience, they must also differentiate themselves from the reductive physicalist by accepting something the reductionist cannot.

---

a priori and that Global Supervenience is a form of what Chalmers (1996) calls "logical supervenience".

The obvious candidate is one or the other restricted version of Nonreductionism introduced in the previous section. The *reductive physicalist* – theorists like Smart (1959), Lewis (1966, 1972) and Armstrong (1981) – cannot accept that mental phenomena are weakly modally distinct from neurophysiological phenomena. If mental phenomena *m* is identical with neurophysiological phenomena *n*, then, according to (I), *m* and *n* necessitate one another. The weak modal distinctness of *m* and *n* means that one or another of these necessitations does not hold. Given that strong modal distinctness entails weak modal distinctness (see fn. 15), the assumption of Weak Restricted Nonreductionism is the thesis that differentiates the reductive from the nonreductive physicalist. This difference is captured in the following table:

	Weak Restricted Nonreductionism
Reductive Physicalist	(-) Mental and neurophysiological phenomena symmetrically necessitate one another
Nonreductive Physicalist	(+) Mental and neurophysiological phenomena are *at least* weakly modally distinct

Table 2: Reductionism and Nonreductionism

Just as there are distinctions that can be made between versions of physicalism, the ambiguity in the assumption of Nonreductionism shows that there are distinctions that can be made within nonreductive physicalism.

The candidate that differentiates versions of nonreductive physicalism is the assumption of Strong Restricted Nonreductionism. A nonreductive physicalist

might find fault with the modal arguments developed by Kripke (1980), but still accept the argument from multiple realizability. Plausibly, this version of nonreductive physicalism, which we might call L-nonreductionism, maintains that mental phenomena do not necessitate neurophysiological phenomena, but the reverse necessitation holds.[18] In other words, L-nonreductionists accept a *local supervenience* thesis of mental phenomena on neurophysiological phenomena:

> (M-N Local Supervenience) Any minimal neurophysiological duplicate of an actual individual is a psychological duplicate *simpliciter* of that actual individual.

Mental phenomena, though weakly modally distinct from neurophysiological phenomena, are asymmetrically necessitated by them. Roughly, this supervenience thesis captures the position of many individualists about intentional mental properties, namely that Putnamian Twins *are* psychological duplicates (see Section 1.4 for more discussion on individualism and anti-individualism).

However, nothing about the general thesis of physicalism requires that one accept M-N Local Supervenience. All the nonreductive physicalist must maintain is that the *physical* way our *world* is, not the *neurophysiological* way it is, metaphysically necessitates the psychological way it is. Recall, if physicalism is true, then 'if P then M' is a necessary truth and P reports all of the physical facts; not just the neurophysiological facts. Therefore, it is consistent with one version of nonreductive physicalism that it is metaphysically possible for two *individuals* to

[18] If we construe the neurophysiological phenomena to be properties of the whole brain or neural events occurring "holistically" rather than in local regions of the brain, Yablo (1992b) seems to be an example of a nonreductive physicalist who accepts this *at least with respect to some mental phenomena* (e.g., phenomenal properties and events and/or intentional phenomena with narrow content). See Section 5.2.1.

be neurophysiological duplicates without being psychological duplicates.[19] Plausibly, this version of nonreductionism, which we can call G-nonreductionism, accepts the entirety of Kripke's modal arguments and/or finds anti-individualist considerations about mental content persuasive. In short, G-nonreductionists consider mental phenomena to be strongly modally distinct from neurophysiological phenomena. The differences between L- and G-nonreductionists are captured in the following table:

	Strong Restricted Nonreductionism
L-Nonreductionism	(-) Mental phenomena are *weakly modally distinct* from neurophysiological phenomena but the latter necessitates the former
G-Nonreductionism	(+) Mental phenomena are *strongly modally distinct* from neurophysiological phenomena

Table 3: Two Kinds of Nonreductionism

One might be inclined to assimilate G-nonreductionism with nonphysicalism, since the latter view is commonly understood to posit only a *contingent* connection between mental and neurophysiological phenomena.[20] This is precisely the kind of connection the G-nonreductionist takes to hold between mental and neurophysiological phenomena, since they accept Strong Restricted

[19] Think here of the twins in Putnam's Twin Earth cases. *Ex hypothesi* they are molecule for molecule identical to one another and so neurophysiologically exactly the same, but their psychologies differ, since they instantiate different intentional mental properties. The possibility of Putnamian Twins is not taken to undermine physicalism and historically is one reason for formulating it in terms of a global rather than local supervenience thesis.

[20] In what follows, I assume that G-nonreductionism involves a commitment to physicalism.

Nonreductionism. The assimilation, however, is a mistake. First, what defines nonphysicalism is the rejection of Global Supervenience, not their acceptance of some version of Nonreductionism. By accepting that there are minimal physical duplicates of the actual world that are not psychological duplicates of the actual world, the nonphysicalist has a substantive disagreement with all physicalists including the G-nonreductionist. G-nonreductionists accept, while the nonphysicalist denies, that 'if P then M' is a *necessary truth*. Second, most versions of nonphysicalism accept Strong Restricted Nonreductionism like the G-nonreductionist. In other words, if N_i reports a subset of the facts reported by P, namely all the neurophysiological facts concerning a specific individual, and M_i reports all of the mental facts of that individual, then both the G-nonreductionist and the nonphysicalist hold that 'if N_i then M_i' is a *contingent truth*. But an important difference remains even about the status of this truth. In order to see this difference, consider an important argument set out by Terrance Horgan (1993).

Horgan (1993) has forcefully argued that all versions of physicalism – reductive and nonreductive alike – cannot take the *metaphysically necessary* connections between the mental and the physical as *explanatorily brute*. He claims there must be *in principle* some deeper and physicalistically acceptable explanation as to why there is this necessitation from the physical to the mental.[21] In other words, despite being necessary, the physicalist cannot hold that 'if P then M' is a brute fact about the actual world. The nonphysicalist is in a decidedly different

[21] What Horgan thinks the physicalist owes us is a kind of *non-causal* explanation of supervenience.

position. They contend that 'if P then M' is contingent and, furthermore, that the search for why this contingent connection holds is misguided. For example, things just happen in accordance with the fundamental physical laws. Likewise, things just happen in accordance with these fundamental psychophysical laws and "there is no asking 'how'" (Chalmers 1996, 170). The proposition 'if P then M' is an explanatorily brute contingent fact about the actual world.

Given that G-nonreductionists are physicalists, I think we should extend the arguments of Horgan (1993) and apply them to the *metaphysically contingent* connections they posit between mental and neurophysiological phenomena. In short, G-nonreductionists cannot accept that 'if N_i then M_i' is a *brute* contingent fact about our world. Other contingent truths are explainable and the G-nonreductionist should think that this contingent truth is as well. This extension of Horgan's argument is vindicated by the fact that most nonreductive physicalists nowadays take the mental to be *realized in* the neurophysiological. The nonphysicalist and the nonreductionist can both consistently hold Strong Restricted Nonreductionism. However, the nonphysicalist claims that 'if N_i then M_i' is a brute contingent fact, while the G-nonreductionist holds that it is explainable in terms of a *realization relation* between the mental and the neurophysiological. In addition to the acceptance of Global Supervenience, this is the defining difference between G-nonreductionism and nonphysicalism.

What, then, is the answer to the question posed at the end of Section 1.3: Which restricted version of Nonreductionism must the nonreductive physicalist

accept? The nonreductive physicalist differentiates themselves from the reductive physicalist by accepting Weak Restricted Nonreductionism, so they must accept *at least* this. I have argued, however, that Strong Restricted Nonreductionism is consistent with Global Supervenience so there is a version of nonreductive physicalism that accepts something stronger than Weak Restricted Nonreductionism. The ambiguity in Nonreductionism infects the position of nonreductive physicalism, splintering it into two positions: L- and G-nonreductionism. In later sections we will see that this ambiguity creates similar problems for some of the other assumptions taken to generate the exclusion problem. The way I elect to proceed in light of this ambiguity is to define a notion of modal distinctness that is general enough to subsume both weak and strong modal distinctness. Here is the obvious proposal:

> (Modal Distinctness) x is *modally distinct* from y just in case it is possible that x is present and y absent or it is possible that y is present and x absent.

where the 'or' is to be read as *inclusive* disjunction. Equivalently, x is modally distinct from y just in case either (i) the (a)-condition in (I) fails to be satisfied, (ii) the (b)-condition in (I) fails to be satisfied, or (iii) both conditions in (I) fail to be satisfied. Both the L- and G-nonreductionist can agree that mental and neurophysiological phenomena are modally distinct, hence, Nonreductionism can be *re*formulated as follows:

> (Nonreductionism) For every neurophysiological phenomenon *n* and every mental phenomenon *m*, *n* is modally distinct from *m*.

Again, anyone who accepts either the weak or strong versions of Nonreductionism can accept this reformulation. Therefore, this essay will consider the exclusion problem as being generated by *inter alia* the assumption of Nonreductionism as just stated. Furthermore, it will examine a resolution to this problem from the point of view of a nonreductive physicalist, where this entails a commitment to Global Supervenience and Nonreductionism (restricted and clarified as above). Therefore, the reductive solutions advanced by authors such as Kim (1993b, 1998, 2005) will be of only secondary concern.

Section 1.4: Two Problems of Mental Causation

Making sense of mental causation faces at least three different problems.[22] In this section, I will discuss two contemporary problems of mental causation that share an important similarity to the problem that forcefully presented itself to Descartes. Roughly, these problems are making sense of how *anomalous* mental phenomena and *extrinsic* mental phenomena can be causal. The issue, as the discussion will hopefully make clear, is that if the mental is either of these ways, then its very nature threatens to preclude its causal efficacy. This is in stark contrast to the problem of mental causation presented in the following section in which it is *assumed* the mental has the potential to causally influence the physical.

The first of these problems arises most forcefully for those who endorse Donald Davidson's (1980d) anomalous monism. Davidson argues for the thesis he calls "anomalism of the mental" which states that there are no *strict* laws on the

[22] See (Kim 1998) for these three different problems. My discussion in this and the next section is heavily influenced by Kim's discussion and analysis.

basis of which mental events can be predicted or explained or can predict or explain other events. Exactly how Davidson argues for this claim is a matter of considerable controversy, but what has been clear for much of the philosophical community is that anomalism threatens to lead to what Brian McLaughlin (1989) has labeled "type-epiphenomenalism".[23] The problem is to make sense of how *anomalous mental properties* can be causal properties. More precisely, if mental properties are anomalous, then nothing is efficacious *in virtue of* possessing these properties. The thesis of "anomalism of the mental" appears to threaten the causal efficacy of mental properties.

According to Davidson, mental events enter into causal relations with physical events. Furthermore, he endorses the view he calls the "nomological character of causality", that is, every instance of causation is subsumable by a strict law. But since there are no strict psychophysical laws[24], mental events that enter into causal relations with physical events must be subsumed by some strict physical law. Events are subsumed by laws, we may suppose, in virtue of their properties. So the mental events that enter into causal relations with physical events must have physical properties and hence be physical events. The conclusion Davidson draws is that commonsense facts about mental causation, the nomological character of causality, and anomalism of the mental lead to a token-identity theory which holds that every mental event that enters into causal relations with physical events is itself a physical event.

---

[23] Davidson himself, not surprisingly, denies that it has this consequence (see (Davidson 1993)).
[24] This presumably follows from the stronger thesis of anomalism of the mental.

No matter how congenial this picture of the mental looks to mental causation, epiphenomenalist worries loom large. Ernest Sosa expresses this worry,

> The being of a desire by my desire has no causal relevance to my extending my hand … if the event that is in fact my desire had not been my desire but had remained a neurological event of a certain sort, then it would have caused my extending my hand just the same (Sosa 1984, 278).

The event that causes the extending of my hand enters into this causal relation in virtue of being subsumed by some strict physical law, and it is subsumed by this law because of the physical properties it possesses. But in no way are the mental properties of this event relevant to it causing anything, since there are no strict psychophysical laws to subsume these events. This is why Sosa complains that the fact that my desire is a desire has no causal relevance to the extension of my hand. The physical properties of my desire look to be the only causally relevant properties it possesses. The problem is that, *in virtue of being anomalous*, mental properties are just not the right kinds of things to make a causal difference. The problem stems from the nature of the mental itself, namely its anomalousness.

The second of these problems arises after one takes seriously Putnam's (1975a) and Burge's (1979) arguments for anti-individualism about mental content. Anti-individualism is a thesis about the kinds of physical properties that do not necessitate intentional mental properties. Specifically, it claims that the localized (i.e., intrinsic) physical properties of cognitive agents do not necessitate intentional mental properties. Anti-individualism can be understood as a rejection of M-N Local Supervenience and, therefore, inconsistent with the position I have labeled L-

reductionism (see Section 1.3.1). The most famous considerations in favor of this view are the reflections on Twin Earth-type scenarios.

Tyler Burge's original thought experiment[25] asks us to imagine a member of our linguistic community, call him Jones, who has many true beliefs about the ailment we call 'arthritis'. Jones believes that having arthritis in the joints is painful, that arthritis is an inflammation of the joints, that there are many different types of arthritis, etc. However, Jones also has some false beliefs about arthritis, in particular that the pain in his thigh is due to arthritis. When Jones informs his doctor about this pain, his doctor tells him that since arthritis can only afflict the joints, the pain in his thigh cannot be due to arthritis. Showing deference to the medical expert, Jones quickly revises his position and no longer believes that he has arthritis in his thigh.

Next we are asked to imagine a microphysical duplicate of Jones, an individual who is molecule for molecule identical and so, we may suppose, neurophysiologically identical to Jones – let's call him Twin-Jones. Twin-Jones is a member of a linguistic community on Twin-Earth, which is almost identical to Earth save that according to its conventions the term 'arthritis' applies to both inflammations of the joints *and ailments in the thigh*. When Twin-Jones informs his doctor about the ailment in his thigh by saying, 'I believe I have arthritis in my thigh,' his doctor does not correct him. Both Jones and Twin-Jones express their initial beliefs with the same sentence-form, but only Twin-Jones's belief is true.

[25] Burge (1986) presents a different version of this argument that does not assume what many take to be a controversial premise in his argument from (1979): that one can possess a concept while only incompletely grasping it.

The preliminary conclusion is that the content of their mental states must differ. But given that they are neurophysiologically exactly alike, it cannot be that their neurophysiological conditions necessitate their intentional mental states. The only difference between Jones and Twin-Jones is that they belong to different linguistic communities. The anti-individualist draws the conclusion that the supervenience base of intentional mental properties must include details about one's neurophysiology supplemented with details about one's linguistic community.

The conception of intentional mental properties as *extrinsic-relational-historical* properties of cognitive agents has arisen out of considerations similar to these. The problem this raises is the tension created between the extrinsicness of mental properties and the philosophically intuitive conception of causation in which the causal properties of an object are its localized (i.e., intrinsic) properties. Since intentional mental properties are not properties like these, they do not look to be good candidates for being among the causes of our bodily movements. A famous passage by Fred Dretske illustrates the problem intentional mental properties appear to face,

> Something possessing content, or having meaning, can be a cause without its possessing that content or having that meaning being at all relevant to its causal powers. A soprano's upper register supplications may shatter glass, but their meaning is irrelevant to their having this effect. Their effect on the glass would be the same if they meant nothing at all or something else entirely different (Dretske 1988, 79).

The presumption is that the soprano's upper register supplications have meaning and it means what it does, not in virtue of its intrinsic character, but rather in virtue

of some of its extrinsic features. These extrinsic properties of the supplications made no difference to whether it shattered the glass. The content properties of the events that cause behavior are supposed to be analogous to the meaning possessed by the supplications – causally irrelevant *given their extrinsic character*.

Coupled with a prevalent and influential cognitive theory that mental processes are computational processes, the problem only seems to be worse. For the computational theory of thinking explicitly endorses the view that the causally relevant properties that drive the computational process are the *syntactic properties* of the representational elements the computations are defined over. When these computational processes are implemented in human neurophysiology, it is the neurophysiological features of a cognitive agent that stand to be best suited for the "syntactic" equivalents that are efficacious in the production of, e.g., bodily movements. The problem is that, in *being extrinsic*, intentional mental properties are just not the right kinds of things to make a causal difference. Like the problem of the anomalousness of mental properties, this problem stems from the very nature of the mental itself. Intentional mental properties being extrinsic rules out their being causally relevant.

Both of these problems of mental causation have the same general structure: mental properties have feature F essentially; properties with feature F cannot be causally efficacious properties; therefore, mental properties cannot be causally efficacious properties. And notice the similar reasoning involved as to why mental properties face a problem when it comes to causation. Sosa claims that the event

which causes the extension of my hand is not efficacious *qua* mental (viz., being a desire to extend my hand), since had the event not been a desire the effect would have occurred all the same. Similarly, Dretske says that the event which causes the shattering of the glass is not efficacious *qua* event with a certain content, since had the event not had this content the effect would have occurred all the same. Presumably, intentional properties face the same problem when considerations about causation and computation enter the picture – had the event that caused my bodily movement not had these intentional mental properties, the bodily movement would have occurred just the same. *Prima facie*, the *anomalousness* and *extrinsicness* of mental properties precludes their being causally efficacious properties.[26]

Section 1.5: The Exclusion Problem

The third problem of mental causation – the problem that will be the focus of this essay – is known as *the exclusion problem* and differs from the problem of anomalousness and extrinsicness. The source of this problem does not lie with the nature of the mental but rather with a feature of the physical world. Bennett (2003) nicely points out the difference here,

> Those worries turn on claims about the failings of the mental – that it is not spatially extended, or is not invoked in the requisite sort of strict laws, or is somehow inappropriately extrinsic. The exclusion problem, in contrast, does not purport to show that mental events and properties are somehow by their nature unsuited to causing anything. It is rather that even if they are perfectly suited to causing

[26] Notice that these arguments only work if the event that caused my bodily movements can occur without it possessing any (intentional) mental property. In other words, these considerations assume that intentional mental properties are not essential to the events that cause bodily movements.

> things, there is nothing around for them to cause (Bennett 2003, 471).

The anomalousness and extrinsicness of the mental looks to undermine the very possibility of mental causation. On the contrary, what Bennett claims is that we can assume mental phenomena to be, by their very nature, fully capable of causing things; the possibility of mental causation is not what is at issue. Instead, the exclusion problem forces one to make sense of how mental phenomena *actually* cause things if the physical is "sufficient" and accomplishes all of the causing that needs to be done.

Section 1.5.1: The Principle of Completeness

The idea of the physical doing all of the causal work brings to mind a world in which, were we capable, tracing the entire causal history of a physical phenomena would never require us to go beyond the realm of physical phenomena. If a physical phenomenon has a causal history at all, it has a complete physical causal history (Menzies 2003, 197). According to David Papineau, this idea of the physical world as being *causally complete* is "purely a doctrine about the structure of the physical world. It says that if you start with some physical effect, then you will never have to leave the realm of the physical to find a fully sufficient cause for that effect" (Papineau 2001, 3). Following Kim (2005), the idea of the causal completeness of the physical world can be codified in the following principle:

> (Completeness) For every physical phenomenon p that has a sufficient cause occurring at t, some physical phenomenon p^* is causally sufficient for p at t.

This formulation presupposes that the physical world is deterministic and some may take this to be problematic. However, this is of little consequence, since if indeterminism is true, Completeness may be reformulated as follows: for every physical phenomenon *p* that has its probability distribution fixed at t, some physical phenomenon *p** fixes this probability distribution at t. Since it is not obvious that anything concerning mental causation hangs on the truth of determinism or indeterminism, I shall work only with the deterministic formulation. Before moving on it would be good to spend a little time getting clear on the assumption of Completeness.

First, Completeness should be distinguished from the stronger assumption that the physical world is *causally closed*:

> (Closure) Any phenomena causally sufficient for some physical phenomena is itself a physical phenomena.

Notice how Papineau informally characterizes the completeness assumption by saying that you *never have* to leave the physical realm to find a fully sufficient cause. Closure, on the other hand, states that you simply *cannot* leave the physical realm to find a fully sufficient cause. It follows from Closure, but not Completeness, that physical effects do not have sufficient nonphysical causes.

In addition, it was causal completeness, not causal closure, that initially generated worries about the mental being causally excluded. Norman Malcolm introduces the problem,

> Given the antecedent neurological states of his bodily system together with the general laws correlating those states with the contractions of muscles and movements of limbs, he would have

> moved as he did regardless of his desire or intention. If every movement of his was completely accounted for by his antecedent neurophysiological states (his "programming"), then it was not true that those movements occurred *because* he wanted or intended to get his hat (original emphasis) (Malcolm 1968, 53).

Suppose we were capable of tracing the complete causal history of some bodily movement. The suggestion is that what we would find is a causal chain of neurophysiological states leading up to the bodily movements in question. Malcolm's worry is that if these physical states of the body and brain provide a complete causal history of our bodily movements, then every bodily movement will have a sufficient neurophysiological cause; the presence of mental states would not have made any difference to the occurrence of these bodily movements. The problem is not that neurophysiological states are the *only* states that can be causally sufficient for bodily movements. Rather, it is that the causal work the mental is assumed to do *vis-à-vis* bodily movements is already done by these physical states. The problem, at least as Malcolm saw it, is that our neurophysiology provides a causally complete account of our bodily movements, not that it provides a causally closed account of such movements.[27]

What I have claimed differentiates Completeness from Closure is that the latter, but not the former, rules out physical effects from having nonphysical causes. Giving this claim substance requires that we face the terminological issue of

[27] What I am calling a causally complete account of our bodily movements Malcolm refers to as "mechanism". He says, "The version of mechanism I wish to study assumes a neurophysiological theory which is adequate to explain and predict all movements of human bodies except those caused by outside forces" (Malcolm 1968, 45). Claiming that a theory is adequate to explain and predict some phenomena is consistent with it not being the only theory adequate to explain and predict some phenomena (see (Kim 1989) for issues concerning multiple explanations). The problem begins with the completeness of neurophysiology, not with its closure.

specifying what is meant by 'physical'. This is a controversial matter and I presently have nothing new to add to this debate. I do think it is useful for our present purposes, however, to follow Papineau (2001) in stipulating that 'physical' means *non-mental*. His justification for this usage is that this sense of the term fits nicely with the line of reasoning that has come to be called the "causal argument for physicalism": for every physical phenomena *p* that has a sufficient cause occurring at t, some physical phenomena *p** is causally sufficient for *p* at t; all mental phenomena have physical effects; the physical effects of mental phenomena are not overdetermined; therefore, mental phenomena must be identical with physical phenomena. Various kinds of causal arguments have been advanced in support of the identity of mental and physical phenomena and these arguments are both philosophically interesting and controversial. If 'physical' meaning non-mental allows these arguments to remain interesting and controversial, this is *prima facie* reason to continue with such a usage of the term.

Adopting this usage allows 'physical' to be construed broadly enough to include many of the common and everyday phenomena we intuitively take to be physical (e.g., tables, chairs, cars, pool sticks, pool balls, bodily movements, etc.) and, furthermore, it allows us to include all of the phenomena cited in or subsumed by the lower-level sciences to count as physical (e.g., neurophysiological, biological, chemical, and microphysical phenomena). What it does rule out are the intentional (e.g., beliefs, desires, intentions, emotions, etc.) and phenomenal phenomena (e.g., pains, sensations, feelings, etc.) cited in our psychological (both

folk and scientific) explanations of purposeful behavior. This is surely a virtue since no intuitive categorization of these phenomena would include them as physical.

Additionally, multiple realizability and modal arguments have made it almost philosophical orthodoxy to reject the view that mental phenomena are neurophysiological phenomena. The philosophical response has been to claim that though distinct from such physical phenomena the mental is still realized in the physical. The prevalence of such phrases in discussions about the mind suggests a conception of the physical that includes the brain and all of its neurophysiological properties as physical phenomena. Capturing both our intuitive and philosophical categorizations in this way is more *prima facie* reason to adopt the sense of 'physical' as non-mental.

Therefore, for the remainder of the present essay, I shall adopt the sense of 'physical' as *non-mental*, counting both macroscopic and microscopic phenomena alike as physical. The principle of Completeness, then, takes on the following interpretation: for every non-mental phenomenon y that has a sufficient cause occurring at t, some non-mental phenomenon x is causally sufficient for y at t. A consequence of adopting this interpretation is that Completeness holds in both the micro-world of chemical, atomic, and quantum phenomena and in the macro-world of car accidents, tornadoes, and cab hailings (*contra* Baker (1993) and Sturgeon (1998)). In addition, the commitment of the nonreductive physicalist to Global Supervenience should be understood in a way continuous with this interpretation of

Completeness: the *non-mental* way our world is metaphysically necessitates the mental way it is. If God were to fix the non-mental (i.e., non-intentional and non-phenomenal) way our world is, no more work would need to be done in order for him to fix the mental way our world is.

Let it be noted, however, that typically in the context of discussing mental causation the completeness of the physical world is given a weaker interpretation (Menzies 2003, 197). We are asked to trace the causal history of not just any phenomena, but, specifically, bodily movements. Furthermore, the idea is that in tracing the causal history of these phenomena we will find a chain of neurophysiological phenomena leading up to the bodily movements in question. Codified into a principle we can say that for every bodily movement *b* that has a sufficient cause occurring at t, some neurophysiological phenomenon *n* is causally sufficient for *b* at t.[28] It seems to me that there is no need to insist on the causal completeness of the physical world *as a whole* to generate problems for mental causation. Instead, all we need to insist on is what an optimistic neuroscientist would insist on, namely that neurophysiology can provide a complete causal history of all the intentional movements of our bodies. For the most part, this weaker interpretation of Completeness will be the assumption at issue.

Section 1.5.2: The Principle of Exclusion

[28] The bodily movements I am concerned with are endogenously produced bodily movements. If Jones lifts my arm, then my body has moved. But this kind of bodily movement is not considered to be the effect of either the mental ways I am nor the neurophysiological ways I am (unless of course I told Jones to raise my arm).

If physical phenomena like bodily movements have sufficient causes at all, then Completeness entails that they have sufficient physical causes. On rather ordinary empirical assumptions (and along the lines of the weaker interpretation typically given to Completeness in the context of discussing mental causation), the sorts of physical phenomena that cause bodily movements are neurophysiological. It follows that neurophysiological phenomena do all of the causal work there is to do *vis-à-vis* bodily movements. The result is that there looks to be no causal work left for mental phenomena to do *vis-à-vis* these movements. As Malcolm worried: *wouldn't my body move just as it actually does regardless of my actual mental condition*? A simple solution to this problem is to accept the view that mental phenomena just are neurophysiological phenomena. This mental-neurophysiological identity promises to completely bypass any worries about causal exclusion for

> On the identity view, there is here one cause ... not two. As for explanation, at least in the objective sense, there is one explanation here, and not two. The two explanations differ only in the linguistic apparatus used in referring to, or picking out, the conditions and events that do the explaining … they both point to one objective causal connection, and are grounded in this single causal fact (Kim 1989, 87).

This is the point at which the exclusion problem becomes most troublesome for the nonreductionist, since on their view mental phenomena are *modally distinct* from neurophysiological phenomena.[29] Of course, the reductionist is not completely out

[29] Identifying mental phenomena with physical phenomena does not, without some further assumptions, avoid the exclusion problem. For, if mental phenomena turn out to be non-neurophysiological physical phenomena they still stand to be causally excluded by

of the water either, since multiple realizability and modal arguments are strong reasons in favor of the modal distinctness of the mental and the neurophysiological. Either one embraces Nonreductionism and faces the bugbear of exclusion or one endorses some type of reductionism putting oneself in the tough dialectical position of having to deny the multiple realizability of the mental and rejecting the modal arguments for distinctness. Whatever route one finds most reason to take, the philosophical work is not insignificant.

Upon closer inspection, however, Nonreductionism and Completeness are not enough to generate the exclusion problem all by themselves. Completeness tells us that whatever causal work there needs to be done *vis-à-vis* bodily movements is done by the neurophysiological, and Nonreductionism assures us that none of *that* work is work done by the mental. But this is consistent with the mental's causal work being *redundant* or *superfluous*. In other words, if we are willing to countenance multiple sufficient causes of bodily movements, mental phenomena may still be causally relevant despite the complete account of these movements being given by the neurophysiological. Perhaps, this line of thought goes, my mental condition does not matter to the movement of my body, since it would have moved either way. But this is consistent with my mental condition being something extra, a gratuitous cause of my bodily movements but a cause nonetheless.

---

neurophysiological phenomena. The problem is not that the mental isn't physical but that the mental isn't *neural*.

However, the following consideration is highly plausible: unless these bodily movements are *genuine cases of overdetermination* then the neurophysiological simply excludes the mental from causing them. Kim (2005) takes considerations like this to involve the application of an *exclusion principle*: "No single phenomenon can have more than one sufficient cause occurring at any given time – unless it is a genuine case of causal overdetermination" (Kim 2005, 42). There are a variety of ways to articulate the principle appealed to here, but for present purposes I will follow Kim (2005) in his general formulation:

> (Exclusion$_{GEN}$) Necessarily, if some phenomenon x is causally sufficient for an effect y, then unless y is overdetermined no phenomenon x* distinct from and existing simultaneously with x is such that it is causally sufficient for y.

A few clarifications are in order.

First, besides its initial plausibility one might wonder if there are any other reasons to accept Exclusion$_{GEN}$. Menzies (2003) interprets Kim as giving the following argument in its favor. According to Kim, the supervenience relation between the mental and the neurophysiological is best explained by taking mental properties to be higher-order functional properties realized by lower-order neurophysiological properties.[30] For some object to possess a higher-order

[30] One might construe Menzies as interpreting Kim as holding M-N Local Supervenience. If so then it should be noted that this is a rather strong form of supervenience and its denial is consistent with Global Supervenience. On the other hand, perhaps Menzies is interpreting Kim only as holding a supervenience thesis that is restricted to a set of worlds less inclusive than the set of metaphysically possible worlds. For instance, the quantifier expressed by 'any' in 'any minimal neurophysiological duplicate of an actual individual is a psychological duplicate *simpliciter* of that actual individual' ranges only over the nomologically possible individuals. If so then it should be noted that this is a rather weak form of supervenience and its acceptance is consistent with the denial of Global Supervenience. If this supervenience thesis is suppose to signal a commitment of the physicalist then it is either too strong or too weak.

property is for it to possess some lower-order property that satisfies a certain condition (Kim 1998, 19). In the case of mental properties, they are higher-order properties because for an object to possess them requires that it possess some lower-order property that has a certain kind of causal profile. According to this characterization, mental property M is the property of having some property P that typically has such and such causes and such and such effects.[31] Whatever properties typically have these causes and effects are said to be realizers of the higher-order mental property.

Furthermore, Kim thinks that the following principle is highly plausible for higher-order functional properties and their realizers:

> (Inheritance) If a higher-order property P_1 is realized by a lower-order property P_2, then the causal powers of P_1 are identical with the causal powers of P_2 (or at least a subset of them).

A consequence of Inheritance is that the causal powers of a higher-order mental property are the same as the causal powers of the property that realizes it.[32] If we suppose that the mental property M is causally sufficient for some bodily movement B, then the manifestation of the power to bring about B is the manifestation of a power the lower-order property P also possesses. There is, then, only really one manifestation of a causal power and so only one causal connection

---

[31] I think the way to understand a property as having a causal profile is in terms of what *causal generalizations* or *type-causal claims* are true of that property rather than which *token-causal claims* are true of that property. So saying that the mental property pain typically causes wincing is to say that the type-causal claim that pain tends to cause wincing is true. And this type-causal claim could be true in a world even if no instance of pain ever causes anyone to wince.

[32] Inheritance is ambiguous: are the causal powers of the higher-order property *qualitatively identical* with those of the lower-order realizer or are they *numerically identical*? I find it natural to read it as claiming the powers are numerically identical and this is how I interpret Menzies as reading Kim.

to B. Therefore, at least as it pertains to mental properties and their lower-order neurophysiological realizers, there will only ever be a single instance of causation that connects them to effects such as bodily movements. This looks awfully close to maintaining that the mental/neurophysiological causal situation is one in which $\text{Exclusion}_{\text{GEN}}$ is true, namely that the presence of one sufficient cause, and so one causal connection to bodily movements, excludes the presence of others assuming that there is no overdetermination.

Second, many think that, in order to prevent it from being obviously false, $\text{Exclusion}_{\text{GEN}}$ pertains only to phenomenon which *exist at the same time*. Yablo (1992b) provides us with some reason for including it:

> Though there may be irreflexive relations *R* whose relata do contend for causal influence as the principle says, for many *R*s this competition arises only sometimes, and for others it never arises. Ironically, *R* = causation is a case in point. Let x be causally sufficient for y. Then taken at its word, the exclusion principle predicts that y owes nothing to the causal intermediaries by which x brings y about. When *R* is causation's converse, the predication is different but still absurd: events causally antecedent to x can claim no role in y's production (Yablo 1992b, 272).

Lacking this rider, $\text{Exclusion}_{\text{GEN}}$ would rule out both causes further back in the causal chain from x and causes intermediate between x and y as being causally sufficient for y. No one thinks there is any problem with an effect having more than one sufficient cause. Worries about exclusion and overdetermination arise when an effect has multiple sufficient causes *existing at the same time*. Whether or not this rider is entirely necessary is not clear.[33] However, not much harm is done

[33] See (Bennett 2003, 478-479).

in including it especially if it gets us to focus on the kinds of causal situations in which the purported causal competitors exist simultaneously. This is standardly taken to be the case in the mental/neurophysiological causal situation, since the mental is supposed to be realized by the neurophysiological. For the sake of brevity, I will not always be explicit in mentioning this rider, but it will be implicit unless noted otherwise.

Third, the principle of $\text{Exclusion}_{\text{GEN}}$ invokes the notion of *overdetermination*. Besides providing a major obstacle for most extant counterfactual theories of causation, there is no consensus as to what an adequate characterization of overdetermination looks like. If we stick to characterizing it at an intuitive level, it looks to involve "two or more separate and independent causal chains intersecting at a common effect" (Kim 2005, 48). A standard example may help to illustrate this point.[34]

> Assassin and Badgirl simultaneously poison Victim's coffee with identical doses of a lethal poison. Either dose by itself would have sufficed for Victim's death. Victim drinks the coffee and dies. He would have survived if the coffee had not been poisoned.

What looks to be true of these kinds of scenarios is that an effect is brought about by two causes[35] where each was itself sufficient for the effect. Whether this is all there is to overdetermination is not clear, but our intuitive characterization of it at the very least requires the presence of two or more sufficient causes.[36]

---

[34] See (Hitchcock 2007, 522 – 524).

[35] Lewis (1973, fn. 12) disagrees and claims that he has no firm intuitions about whether cases of overdetermination involve multiple *causes*. I think that Lewis's intuitions are simply mistaken on this point.

[36] See (Bennett 2003, 477).

Finally, the principle of Exclusion$_{GEN}$ invokes the notion of *distinctness*. In Section 1.2, I discussed two ways in which phenomena can be distinct: weak and strong modal distinctness. As long as there are multiple well-defined notions that can be expressed by the term 'distinct' then, just like Nonreductionism, the principle of Exclusion$_{GEN}$ is ambiguous. If the principle of Exclusion$_{GEN}$ is to engage with Nonreductionism to generate the exclusion problem, both must employ the same notion of distinctness. I elected to proceed by defining a notion of modal distinctness that both versions of nonreductive physicalism could accept and then I reformulated Nonreductionism in terms of this (see Section 1.3.1). I elect to proceed in the same way with Exclusion$_{GEN}$, reformulating it in terms of modal distinctness:

> (Exclusion$_{GEN}$) Necessarily, if some phenomenon x is causally sufficient for an effect y, then unless y is overdetermined no phenomenon x* modally distinct from and existing simultaneously with x is such that it is causally sufficient for y.

To repeat, saying that x* is modally distinct from x is saying that either (i) x fails to necessitate x*, (ii) x* fails to necessitate x, or (iii) neither x nor x* necessitates the other. The idea behind Exclusion$_{GEN}$, then, is that modal differences between phenomena will manifest in the realm of causation. If these phenomena exist simultaneously, then either the phenomena overdetermine their purported effect or one causally excludes the other. At least part of the curiosity of the exclusion problem is whether the modal differences between mental and neurophysiological phenomena manifest in either of these ways.

Section 1.5.3: The Exclusion Dilemma

Now that some of the details of Exclusion$_{GEN}$ have been covered, we can apply it to common causal situations involving mental phenomena to generate the exclusion problem. Completeness tells us that in tracing the causal history of our bodily movements we will find a chain of neurophysiological phenomena that serve as causally sufficient conditions of those movements. Hence, every bodily movement has a sufficient neurophysiological cause. Exclusion$_{GEN}$ says that any effect that has a sufficient cause has no other modally distinct sufficient causes unless it is overdetermined. Let us formulate the relevant substitution instance of Exclusion$_{GEN}$ (which I will simply call 'Exclusion'):

> (Exclusion) Necessarily, if some neurophysiological phenomenon *n* is causally sufficient for bodily movement *b*, then unless *b* is overdetermined no mental phenomenon *m* modally distinct from and existing simultaneously with *n* is such that it is causally sufficient for *b*.

Assuming the truth of Exclusion and Completeness, the *exclusion dilemma* follows: no mental phenomenon *m* modally distinct from and existing simultaneously with *n* is such that it is causally sufficient for *b* unless *b* is overdetermined. Are mental phenomena modally distinct from neurophysiological phenomena? So long as one or the other version of Nonreductionism is true, the answer is 'yes'. The assumption of Nonreductionism says that mental phenomena are modally distinct (either weakly or strongly) from neurophysiological phenomena. Therefore, in conjunction with the exclusion dilemma, it follows that *m* is not causally sufficient for *b* – it is causally excluded by *n* – or bodily movement *b* is overdetermined.

Either the neurophysiological excludes the mental from having bodily effects or mental causes are redundant, overdetermining causes.

Prima facie, neither option is appealing. Many find there to be something metaphysically absurd about the kind of systematic causal overdetermination that the second horn appears to force on us (Kim 1993c, 281). Others merely maintain that "we ordinarily take [overdetermination] to be false, and it is not clear why we should change the belief" (Peacocke 1979, 143). It is, as Stephen Schiffer has said, "hard to believe that God is such a bad engineer" (Schiffer 1987, 148). On the other hand, if the mental is causally excluded by the neurophysiological, we are forced to the conclusion that the mental is entirely epiphenomenal with respect the movement of our bodies. This conclusion requires a radical revision of what we commonly suppose is the case. Some of the most central assumptions we make about ourselves – e.g., that we are morally responsible agents who at times exercise control over what do – are standardly understood to require mental causation. Furthermore, if we can extend the above argument to effects other than bodily movements, say to other mental phenomena, then the epiphenomenalist horn places in jeopardy not just agency, but perhaps our ability to reason and make inferences as well. The exclusion problem is the dilemma of *mental epiphenomenalism* or *causal overdetermination*.

The centrality of mental causation to the manifest image we have of ourselves is to my mind the primary reason to explore the exclusion problem. Perhaps another reason is that this problem does not rely upon assumptions about

how the mental is – e.g., that it is non-extended, anomalous, or highly extrinsic – but instead only on general principles about causation ($Exclusion_{GEN}$), well-grounded doctrines about the causal structure of the human organism (Completeness), and a philosophically respectable position about the relation between the mind and the brain (Nonreductionism). The exclusion problem truly is, as Kim (2005) has said, a "world-knot".

Chapter 2

Resolutions of the Exclusion Problem

There are a variety of ways to respond to the exclusion dilemma and in this chapter I will present some of them. A few warnings are in order before doing so. First, the following is not intended to be exhaustive. What I have included are those views well represented in the literature and those I find philosophically interesting and worthy of further development. Second, the literature on the exclusion problem is enormous and a diverse range of solutions have been developed quite extensively. I will not offer knock-down arguments to any of these strategies. I cannot possibly give any of these views what they are rightfully due in the little space I have dedicated to them. Instead, my brief discussion of them will hopefully make clear which strategies I will *not* be pursuing in the rest of this essay.

To summarize: (a) I am a "strong" realist about intentional and phenomenal mental predicates; (b) I hold that mental and physical causation is a homogenous relation; (c) I maintain an amended version of Completeness, one which states that every bodily effect that has a cause has a neurophysiological cause; (d) I accept that mental and neurophysiological causes have the same bodily effects and that there is sufficient motivation to reject a token-identity theory (viz., mental *events* = neurophysiological *events*); and, finally, (e) I reject a type-identity theory (viz., mental *properties* = neurophysiological *properties*). My overall philosophical perspective is the mainstream view of contemporary analytic philosophy, that of a

nonreductive physicalist, and my preferred solution to the exclusion problem will reflect this perspective within the confines of the theses (a) – (e).

Section 2.1: Weak Realism

John Heil (1999) has argued that most of the recent attempts at understanding multiple realizability leave us with an obscure ontological picture of the relation between higher- and lower-levels of reality. Specifically, the functionalist inspired attempt to understand the relations that hold between levels of reality as realization engenders more problems than it solves. There is no doubt that our descriptions of the world can be hierarchically ordered, but the mistake is in "reifying the hierarchy, imagining that it corresponds to ontological strata" (Heil 1999, 204). He characterizes the problem as having its source in the dogma that realism about a given set of predicates requires that those predicates to designate properties (Heil 1999, 199).[37] Heil's proposal is to reject this realist dogma and show that doing so has many welcome consequences. One of the more significant consequences is that worries about causal exclusion are dissolved.

The view put forth by Heil is one in which,

> Many predicates apply to objects in virtue of properties possessed by those objects. Of these predicates, some designate properties shared by objects to which they apply. Others do not. Realism about a given predicate, 'ϕ', requires only that 'ϕ' applies truly to objects in virtue of properties actually possessed by those objects.

[37] Heil (2003) identifies the problem more broadly as the tacit philosophical acceptance of the Picture Theory of Language. This family of theories about the way in which representations, including linguistic representations, relate to the world they represent involves the commitment to the realist dogma mentioned above. As far as I can tell, Heil offers no straightforward argument against this Picture Theory of Language and its corollaries concerning realism. Instead, he offers an alternative theory (part of which involves what I've dubbed 'weak realism') and accepts it on the basis of its power to "resolve a wide range of problems in a natural way" (Heil 2003, 14).

> Realism does not require that 'ϕ' designate a property shared by every object to which it truly applies. If 'ϕ' does designate a property, then objects satisfying 'ϕ' must be identical (or exactly similar) in some respect, a respect in virtue of which 'ϕ' holds true of them. The bulk of our predicates, however, are satisfied by ranges of similar, but not exactly similar properties (Heil 1999, 200 – 201).

If we concern ourselves specifically with the mental predicates employed in vernacular psychological discourse, Heil's position is that mental predicates do not designate properties, but nevertheless apply truly to objects in virtue of the (non-mental) properties actually possessed by those objects. In fact, there are no mental properties at all. Heil does not endorse eliminativism about the mental, since, unlike eliminativism, his view admits that mental predicates apply truly to objects. Nor is his view reductionist, since "reductionism requires that there be something to reduce" (Heil 1999, 201). The view is a "weak realism" in which *mental predicates do not designate mental properties* but, all the same, apply truly to objects in virtue of properties those objects actually possess.

The resolution of the exclusion problem is anticlimactic. If there are no mental properties at all, it is not possible for those properties to be causally excluded by neurophysiological properties. As it turns out, our worries about epiphenomenalism and the causal irrelevance of mental properties are perverse. He says, "Your belief, desire, and intention could be epiphenomenal, however, only if they existed a part from your neurophysiological condition" (Heil 2003, 45) and since no mental properties exist at all, *a fortiori*, they do not exist a part from your neurophysiological condition. Additionally, neither is it possible for mental

properties to be causally redundant and overdetermine their effects, since there are no such properties. One might suppose that Heil's view commits him to the troubling claim that it is false that your mental condition causes you to do anything. His response to this is interesting,

> On this occasion, it is true in virtue of your being in [neurophysiological] state P_1 that you have these beliefs and desires, and it is true, by virtue of your being in [neurophysiological state] P_2, that you have this intention. It is true, as well, that your belief and desire caused you to form the intention, true in virtue of P_1's causing P_2 (Heil 2003, 45).[38]

Even if there are no mental properties designated by the predicates of vernacular psychology, claims concerning mental causation are not systematically false. Instead, their truth is grounded in "causal occurrences involving the truth-makers for [mental] predicates" (Heil 2003, 46).

I find Heil's weak realism persuasive especially when applied to certain kinds of predicates (e.g., moral and ethical ones). Furthermore, it seems difficult to accept his weak realism with respect to one area of discourse, but reject it with respect to another. What reason is there for thinking that, e.g., mental predicates designate properties, but moral and ethical predicates do not? Regardless of these issues, I think weak realism about the mental and, in particular, about mental *causation* is a position of last resort. Much of the motivation for accepting weak realism about the predicates of vernacular psychology is that multiple realizability

[38] This sounds similar to Kim's (1984) proposal of supervenient causation, however, there are important differences. Both Kim and Heil would agree that lower-level facts about physical causation are the truth-makers of claims of mental causation. Where they disagree is that Kim thinks these facts about physical causation make higher-level mental properties causally relevant or efficacious properties while Heil denies that there are any such higher-level mental properties at all.

and the level conception of reality has engendered more problems than it is worth. It would be reasonable to suppose that one of these problems is the exclusion dilemma. My suggestion, then, is to understand the present project as an attempt to gain a dialectical advantage against Heil's weak realism. I want to show that whatever other problems multiple realizability and the level conception of reality engender, it does not pose an irresolvable exclusion dilemma. Hence, for the remainder of this essay, I'll assume a strong realism about mental predicates and consider talk of mental causation as picking out a causal relation involving nonreducible mental phenomena.

Section 2.2: The Homogeneity Assumption

To the best of my knowledge, Tim Crane (1995) was the first to make explicit the Homogeneity Assumption and describes it as such, "The labels 'mental' and 'physical' as applied to causation are really transferred epithets – what is mental and physical are the relata of causation, not the causation itself" (Crane 1995, 219). Hence, one way of avoiding both horns of the exclusion dilemma is to deny that mental and physical causation are *homogeneous relations*. Although Crane (1995) does not appear to endorse this solution himself, he does say that it is one of the more common responses to the exclusion problem (Crane 1995, 232).[39]

More specifically, the idea here is that worries about overdetermination are misplaced if mental causes do not bring about their effects in the same way that

[39] Crane classifies both Dretske (1988) and Yablo (1992b) as denying the Homogeneity Assumption. I classify both of these authors differently. Dretske's distinction between triggering and structuring causes entails that mental and neurophysiological phenomena do not have the same effect and it is this which seems to do most of the work. See Section 2.4. I interpret Yablo, on the other hand, as rejecting a suitably amended version of Completeness. See Section 3.2.

physical causes do. Suppose that mental causes are only *supervenient causes* of their effects along the lines of Kim (1984): mental property M superveniently causes effect e just in case M supervenes on property P and P causes effect e.[40] Alternatively, suppose that mental causes merely *program* for the occurrence of their effects, as Frank Jackson and Phillip Pettit (1990) argue all special science properties do. According to this strategy, mental properties are causally relevant properties, not by being causally efficacious themselves, but instead by "ensuring that there would be some property there to exercise the efficacy required" (Jackson and Pettit 1990, 114).

Both Kim (1984) and Jackson's and Pettit's (1990) strategy of rejecting the Homogeneity Assumption begin by claiming that physical properties are causally relevant for their effects by entering into a *non-derivative* causal relation.[41] Kim says physical properties are *causes of* their effects, while Jackson and Pettit talk about physical properties *being efficacious*. This non-derivative causal relation is not the same type of relation that mental properties enter into with respect to those effects. This is why concerns about overdetermination are simply misplaced. Instead, each defines an alternative notion of relevance in which the non-derivative causal relation enters explicitly. In the end, the mental possesses only a derivative

[40] Kim defines supervenient causation as follows: x superveniently causes y just in case x supervenes on x*, y supervenes on y*, and x* causes y* (where the supervenience in question is strong supervenience). What I have suggested above is not quite this but something very similar. If we used Kim's notion of supervenient causation then mental and physical causes would not have the same effect. Avoiding the exclusion problem in this way has more to do with denying that mental and physical causes have the same effect then it does with denying that mental and physical causation are homogeneous.

[41] Non-derivative is not intended to mean unanalyzable or primitive.

kind of causal relevance – a kind of relevance it has in virtue of *supervening on* or *programing for* the physical. Even though mental properties enter into a derivative causal relation with their effects, this should not condemn *explanations* citing mental causes as defective or incomplete. As Jackson and Pettit emphasize, these kinds of explanations may still provide information, which explanations citing the non-derivative physical causes do not (Jackson and Pettit 1990, 116).

One might complain that although the mental looks to be explanatorily on a par with the physical, metaphysically something has been lost. Admitting that mental causes do not cause in the same way that physical causes do looks to be nothing better than a thinly veiled form of epiphenomenalism (Kim 1998, 74). This appears to be Kim's primary motivation for abandoning his account of supervenient causation. Worries like this might be deflected if we employ a less crude conception of causation.[42] I shall be the first to admit, however, that I do work with a "crude" conception of causation in which mental and physical causation are homogeneous relations. I do not have any knockdown arguments against this strategy, but, again, I consider it a position of last resort. Maybe after all is said and done saving mental causation requires jettisoning our "crude" and homogeneous notion of causation and accepting that it just isn't the same as physical causation. But I think sense can be made of non-derivative mental causation and so I shall take the homogeneity assumption to be true. My disclaimer,

[42] See (Crane 1995, 234) and (Hall 2004a, 2004b).

then, is that the project of the present essay is to vindicate *non-derivative mental causation*.

Section 2.3: Rejecting Completeness

What makes the exclusion problem different from the other problems of mental causation is that its source does not lie with the nature of the mental itself. We can assume that mental phenomena are causal in exactly the same way as physical phenomena. Instead, it is a feature of the physical world – its causal completeness – that threatens mental causation. Therefore, one straightforward resolution of the exclusion problem is to cut it off at its very source by rejecting Completeness.

Lynne Rudder Baker (1993) pursues a resolution of the exclusion problem along these lines, although it is important to note that her discussion focuses on what she calls "closure". Baker's thesis of "closure" states, "for every event that has a physical property … there are sufficient physical conditions for its occurrence and for its having all of its physical properties" (Baker 1993, 78-9). This is very similar to the principle I have called Completeness. However, the major dissimilarity is that Baker understands "closure" to be true only if 'physical' is understood as micro-physical, since "a system is causally closed if and only if the elements of the system interact causally only with other elements of the system" and this only holds for the micro-physical (Baker 1993, 79). However, her objections to "closure" do not require one to understand 'physical' as 'micro-physical'. I believe one could broaden the conception of physical to include macro-

properties of objects and still run Baker's objection. This is what I elect to do in this section. So even though Baker does not concern herself with Completeness, I will formulate her objections using it.

Baker's (1993) argument is that principles like Completeness are part of a metaphysical worldview that "interferes" with "a range of explanations that have been found worthy of acceptance" (Baker 1993, 92). Furthermore, Completeness plays a role in generating problems for macro-causation generally. The exclusion problem would be just one of a variety of ways in which our allegiance to Completeness "subvert[s] our ordinary causal notions … constitutive of law, morality, and everyday life, but also [makes a mockery] of the causal claims and explanations in the special sciences" (Baker 1993, 90). Baker's suggestion is that these consequences are sufficient to motivate a rejection of Completeness.

But the exclusion problem is not just a consequence of Completeness. By itself Completeness neither says nor entails that physical phenomena cannot have non-physical causes.[43] Why reject Completeness and not some other part of the metaphysical worldview that generates the problem (say, Nonreductionism)? Baker admits that the problem is really brought about by the conjunction of strong supervenience and Completeness (Baker 1993, 91). But even this cannot be the whole truth, since Completeness and any single supervenience thesis (whether it be strong or global) is consistent with physical effects having non-physical causes (i.e., one could endorse overdetermination). There must be some special reason to

[43] This is true even if Completeness is read as Baker prefers in which every *microphysical* effect that has a sufficient cause at t has a sufficient *microphysical* cause at t.

isolate Completeness as the main culprit of the metaphysical worldview that generates the exclusion problem. She says, "The trouble caused by the conjunction of [strong supervenience] and [Completeness] can be avoided by rejecting [Completeness] *and rethinking the notion of causation*" (Baker 1993, 91). The problem, as Baker analyzes it, is that Completeness presupposes an erroneous view of causation.[44]

The erroneous view of causation is one "founded on the idea that causation is an 'objective relation' between events". According to Baker, an *objective* causal relation is one the instantiation of which entails nothing about the existence or non-existence of intentional mental states unless intentional mental states serve as the relata of the relation (Baker 1993, 91). She finds this conception of causation erroneous because it fails to count as legitimate many connections we intuitively take to be causal (Baker 1993, 92). She claims that, for instance, Smith's failing French intuitively is among the causes of his ineligibility to play NCAA Division I basketball even though these events "could not occur if there were no intentional psychological states" (Baker 1993, 92).

The replacement she suggests for this objective conception of causation is one in which we "begin with the explanations that earn their keep, rather than with the metaphysics" (Baker 1993, 93). The result is a revised concept of causation

[44] Baker's general strategy for avoiding the exclusion problem is similar to Loewer (2007) who rejects principles like Exclusion for presupposing an erroneous view of causation. Besides the principle they reject, their positions differ in that Baker finds fault with the objective conception of causation supposedly presupposed by Completeness while Loewer finds fault with the productive conception of causation supposedly presupposed by Exclusion. Both of these authors think the problem of mental causation is a problem because our conception of causation is confused.

that becomes primarily an explanatory concept. Armed with this new concept, we can see that the exclusion problem is dissolved,

> We begin with the question: Does what we think ever affect what we do? ... With the reversal of priority of cause and explanation … the original question has an easy answer. For example, when Jill returns to the bookstore to retrieve her keys, what she thinks is that she left her keys on the counter and she wants them back. What she thinks affects what she does in virtue of the following explanatory fact: if she hadn't thought that she had left her keys, then, other things being equal, she wouldn't have returned to the bookstore; and given that she did think she had left her keys, then, other things being equal, her returning was inevitable (Baker 1993, 93).

Baker's overall proposal is that the problems that arise from endorsing a particular metaphysical worldview are sufficient to motivate a rejection of that worldview. Her diagnosis is that the real problem lies in our acceptance of principles like Completeness that presuppose an objective conception of causation. This objective conception is mistaken, since it jeopardizes many connections we intuitively take to be causal. By rejecting Completeness the exclusion problem is dissolved and there is no difficulty in making sense of how what we think affects what we do – these causal facts are grounded in the facts about a whole range of successful commonsense and scientific explanations.

Baker's position is radically revisionary in that causation turns into a fundamentally explanatory and hence epistemic concept rather than a metaphysical one. It is not difficult to see that such a revision of our concept of causation would, in some fashion, vindicate mental causation. I do not think, however, that Baker motivates this approach to the exclusion problem well enough. The crux of her proposal is that the "objective" concept of causation is flawed independently of its

role in generating the exclusion problem, since it fails to accommodate many relations we intuitively take to be causal. In what follows, I suggest that an "objective" concept of causation is not flawed in the way Baker argues, although I readily admit more needs to be said about her overall proposal.

I do not think the example she works with is a causal relation at all because Smith's failing of French is *not a distinct event* from his ineligibility to play NCAA basketball. This relation is more like the relation between Socrates's death and Xanthippe becoming a widow. According to Kim (1993), Socrates's death *constitutively brings about* Xanthippe becoming a widow where constitutively bringing about some event involves *inter alia* bringing it about with absolute simultaneity. Furthermore, the relation here is more intimate than causation, since Socrates's death is *logically sufficient* for Xanthippe becoming a widow (Kim 1993c, 23-4). Similarly, in the above situation, Smith's failing of French causes his ineligibility to play basketball with absolute simultaneity and the former event logically sufficed for the latter event. It is of course true that had Smith not failed French, he would have been eligible to play NCAA basketball, but the truth of this counterfactual does not indicate the presence of causation. The alleged cause event is not distinct in the right way from its purported effect.

Perhaps the example is not the best. Luckily, Baker offers another for our consideration: taking a deduction for the office in your home is among the causes of you being audited by the IRS (Baker 1993, 92). This example does not have the same problems as the previous one. Her issue seems to be with the idea that

causation is an "objective" relation that "exists out there", independent from us. In other words, that causation is a "mind-independent" relation. She cashes the idea of being mind-independent out by saying that, unless the relation holds between mental states, instances of it entail nothing about the existence or non-existence of mental states. The deduction-audit relation, then, is not counted as causal, since there could not be deductions and IRS audits without the existence of some mental states.

It is right to say that the relata here do entail the existence of some mental states. But the fact that the relation entails the existence of mental states does little to suggest that it is not "mind-independent". Even though the deduction-audit relation entails the existence of mental states, we still consider it to "exist out there", independent from us. The reason is because Baker's idea of what "objective" amounts to is far too narrow. The notion of "mind-independence" we should work with is one in which the relation holds independent of the way anyone *theorizes* about it. That is, a relation R is "objective", "exists out there", and is "mind-independent" just in case it holds regardless of us *representing* it as holding or *the way we represent* it as holding. With this broader notion, we can see that the deduction-audit relation does not fail to get counted as "objective" and so still qualifies as causal. No one need represent this relation between the deduction and the audit by the IRS as holding in order for it to hold.

Perhaps an "objective" concept of causation is flawed in some other way and we can avoid the exclusion problem by recognizing this flaw. What has been

suggested above is that Baker's specific attempt at revealing this flaw and thus avoiding exclusion worries is not successful. So, for the remainder of this essay, I will presume that we can make good sense of causation as being "objective" and "mind-independent" (i.e., causal relations hold independent of anyone *representing* them as holding) and, furthermore, that its being so does not present us with an indefensible principle of Completeness. However, later in this essay, I consider a similar though different line of reasoning that attempts to avoid the exclusion dilemma by rejecting an appropriately amended version of Completeness. According to this solution, a proper analysis of causation in terms of proportionality grounds the rejection of the idea that every bodily effect has a *difference-making* neurophysiological cause. The details of this solution are discussed at length in Chapter 3.

Section 2.4: The Dual Explanandum Strategy

The very idea that the neurophysiological threatens to causally exclude the mental presupposes a certain structure of the causal scenario, namely one in which mental and neurophysiological causes are competing to bring about *one and the same effect*. If mental causes do not have the same effects as neurophysiological causes, neither is it possible for the mental to be excluded by the neurophysiological nor is it possible for the mental to be relegated to a mere overdetermining cause by the neurophysiological. The possibility of exclusion and overdetermination are worrisome only if mental and neurophysiological causes are in competition for the same effect. Malcolm's early discussion of the exclusion

problem is sensitive to this, "There *would* be a collision between the two accounts [the mental and the neurophysiological] *if they were offered as explanations of one and the same occurrence of a man's climbing a ladder*" (my emphasis) (Malcolm 1968, 52). Therefore, one solution to the exclusion problem is to simply deny the structure of the causal scenario in which mental phenomena have the same effects as neurophysiological phenomena.

A well-known advocate of this *dual explanandum* strategy is Fred Dretske (1988), who argues that mental events are causes of behavior. However, behavior is not mere bodily movement, but instead *a process* consisting in neurophysiological events standing in causal relations to one another which results in bodily movement. On this view, neurophysiological events are causally sufficient for bodily movements – the end products of behaviors – while intentional mental events are causally sufficient for the behaviors themselves. Dretske is clear that it is a mistake to conflate the process with the end product, so neurophysiological events do not have the same effects as intentional mental events. Although Dretske's distinction between behavior and bodily movement would seem to dissolve worries about exclusion or overdetermination, Marras (1998) has pointed out that Dretske faces a similar exclusion problem for behaviors. The crucial feature is that behaviors and bodily movements are not just type-distinct, but token-distinct as well. Specifically, some token bodily movement is a component of some endogenously initiated process that Dretske identifies with behaviors. So if we consider some token behavior, "the question arises whether

that structured event (or process) is physical or nonphysical. If it is physical, then it ought to admit of a physical (biological) explanation even if it admits of an intentional one, and the exclusion problem arises all over again" (Marras 1998, 450).

Marras pursues an approach similar to Dretske's, but the crucial difference is that the effects of mental causes are type-, but not token-distinct from the effects of neurophysiological causes.[45] According to Marras, on a Davidsonian conception of actions and events, we should view a single token event as falling under both an *action type* and a *movement type*. Additionally, the cause of this event should be understood itself as falling under both an *intentional type* and a *neurophysiological type*. From this picture Marras concludes that "a single event can have both a physical/biological explanation and an intentional explanation: the former explanation explains why a certain type of movement occurred on a certain occasion, the latter explains why a certain type of action occurred on that occasion" (Marras 1998, 450). The idea here is that a cause c *qua* intentional type M explains an effect e *qua* action type A, but that same cause *qua* neurophysiological type N explains only e *qua* movement type M. The exclusion problem for mental and neurophysiological events is avoided, since alleged competing causes end up being the very same event. In other words, mental and neurophysiological causes are

[45] Thomasson (1998) considers an approach very similar to this when she claims that causation occurs "only within a level" such that "there is no upward or downward causation" (Thomasson 1998, 187). The idea here is that mental properties are higher-level properties determined and/or constituted by lower-level neurophysiological properties and, importantly, the effects of higher-level mental properties are only ever other higher-level properties. The mental can be causally relevant in the physical world without fear of being excluded by the physical or fear of overdetermining its effects, since "there is no cross-level causation" (Thomasson 1998, 188) and hence no competition for the same effect.

token-identical. Any remaining issue concerning the causal relevance of mental *properties* is avoided because mental and neurophysiological properties are causally relevant to different effects. Mental properties are relevant to events *qua* action types, while neurophysiological properties are relevant to events *qua* movement types.

One might wonder whether neurophysiological properties are relevant to events *qua* action types, but Marras argues that this is not so. He proposes the following necessary condition for causal relevance

> (Relevance) Where c causes e, where c is F and e is G, c's being F is causally relevant to e's being G only if '~Fc □→ ~Ge' holds

and goes on to claim that the relevant counterfactuals are false (Marras 1998, 448). In other words, for some neurophysiological property N and action type A, it is false that if c were not-N then e would be not-A. He remarks,

> It is false to say … that George would not have intentionally risen from the couch, on that occasion, if the neural event in his brain (which on that occasion was an intention of type M), *had not* been a neural event of type N: for George's intention might have been physically realized by a neural event of type N* (causing him, perhaps, to get up from the couch in a slightly different way) (Marras 1998, 448 – 449).

According to Marras, the worlds relevant to evaluating the counterfactual '~Nc □→ ~Ae' are ~Nc-worlds where N is *replaced* by N*, which also realizes the intentional type M. If Marras intends to be working with the standard Lewis-

Stalnaker semantics for counterfactuals, then I think there is little reason to accept his claim that the counterfactual '~Nc □→ ~Ae' is false.[46]

The task of evaluating counterfactuals with negative antecedents is a tricky business, but at least in causal contexts there is a presumption against so-called "replacement readings" like Marras's. Suppose that Assassin poisons Victim's coffee with substance S (in circumstances where there are no preempted back-up causes), Victim drinks the coffee and dies. We assume that the presence of causation here reveals itself in the truth of the counterfactual (C) 'If Assassin had not poisoned Victim's coffee with S, then Victim would not have died'. However, without a ban on replacement readings, the obviously true counterfactual (C) might turn out false. If we replace Assassin's poisoning of the coffee with substance S with something *closely resembling* this event (e.g., Assassin poisoning Victim's coffee with substance S* where S* is similar to S and also lethal to Victim) then Victim still dies. The dependence we expect between Assassin's poisoning of the coffee with substance S and Victim's death is simply not there.

Generally, replacement readings make counterfactuals like (C) with negative antecedents come out false more often than we would like. In a discussion of fragility, David Lewis remarks,

---

[46] Additionally, I think cases of preemption and trumping preemption provide some reason to doubt Marras's account of causal relevance. Suppose Assassin poisons Victim's coffee with a fast-acting poison and Badgirl poisons the same coffee with a slow-acting poison. Assassin's action is a cause of Victim's death, but Badgirl's action is a preempted cause of the death. Assassin's action *being a fast-acting poisoning* certainly seems to be causally relevant to Victim's death *being a death by poisoning*. But it is false that if Assassin's action had not been of the former type, Victim's death would not have been of the latter type, since Victim's death would still have been a death by poisoning given the presence of Badgirl's slow-acting poison in the coffee.

> What is the closest way to actuality for C not to occur? It is for C to be replaced by a very similar event, one that is almost but not quite C, one that is just barely over the border between versions of C itself and its near alternatives. But … if almost-C occurred instead of C, very likely the effects of almost-C would be almost the same as the effects of C. So our causal counterfactual will not mean what we thought it meant, and it may well not have the truth value we thought it had (Lewis 2000, 190).

Therefore, given the standard Lewis-Stalnaker semantics, the closest ~Nc-world is not a world where N is replaced by N* which also realizes R. Karen Bennett makes this point nicely,

> When you are supposed to imagine c_1 gone, you imagine it gone. You do not worry about how the past would have to be different to make it fail to occur, and you do not worry about what else might occur in its place. You simply snip it away as though you had a metaphysical hole-puncher (Bennett 2003, 482).

Thus, Marras has no guarantee that counterfactuals like '~Nc □→ ~Ae' are false and it remains possible that even this iteration of the dual explanandum strategy faces the exclusion problem all over again. In fact, if we embrace a coarse-grained view of events as Marras does and we interpret Completeness as entailing that every *intentional action* that has a cause has a neurophysiological cause possessing some relevant neurophysiological property N, then we have reason to think counterfactuals like '~Nc □→ ~Ae' do hold.

Marras's solution to the exclusion dilemma is a form of *token-reductionism* where mental and neurophysiological events are identical. This kind of reductionism offers a straightforward solution to the problem posed by overdetermination. But, as mentioned in Section 1.4, the commitment to *type-nonreductionism* leaves him with the horn of (type-) epiphenomenalism, which he

attempts to combat by evoking a partial analysis of causal relevance in terms of counterfactuals. I have tried to give some reason for doubting that the initial attempt to avoid type-epiphenomenalism succeeds, but I nonetheless think Marras's Davidsonian inspired solution deserves serious consideration. However, the present essay pursues an explicitly *non-Davidsonian* solution to the exclusion dilemma, so let me clear about what I think this means.

Although there are good reasons to endorse token-reductionism about events, there are, on the other hand, good reasons to consider *mental and neurophysiological events* distinct. In an attempt to avoid this controversy, I will simply assume throughout the present essay that mental and neurophysiological events are distinct. There is, I believe, *prima facie* justification for this assumption (see, e.g., the modal arguments discussed in Section 1.2.2). Furthermore, I shall also assume that mental and neurophysiological events have the same bodily effects and that the respective properties are relevant to the same things, namely events *qua* action types or, as I have been calling them, bodily movements. What I hope to show in the following chapters is that nonreductionism of either variety (type- or token-) does not lead one into irresolvable difficulties concerning epiphenomenalism or overdetermination.

Section 2.5: Reductionism

The exclusion problem presents us with a scenario in which two distinct phenomena are in competition with one another for causing a single effect. What strikes me as the most obvious solution to the problem is simply to deny that the

phenomena in causal competition are distinct. If there is really only one cause here (and two distinct ways of designating it), then the supposed competition is an illusion. Now, I say that this reductionist solution is the most obvious, but not because it is obviously true. Multiple realizability and modal arguments show *at the very least* that reductionism is not obviously true. Nor do I say that it is the most obvious solution because I think out of all of the assumptions that generate the exclusion problem we have most reason to doubt Nonreductionism. Once again, multiple realizability and modal arguments present forceful reasons in favor of Nonreductionism. Instead, I claim this is the most obvious solution to the exclusion problem because considerations about mental causation have been explicitly used in arguments for a reduction of the mental to the neurophysiological. Not to mention, Malcolm's presentation of the exclusion problem involved quite a bit of effort to exorcize our reductionist sentiments (Malcolm 1968, 53).

There are roughly two traditions of argument for reductionism, both of which base their defense on mental causation. The first originates with the idea that general theoretical considerations speak in favor of a reduction of mental properties to neurophysiological properties. This was argued for and developed in various ways by U.T. Place (1956), Herbert Feigl (1958), and J.J.C. Smart (1959).

According to Place, in order to establish the identity of phenomenal properties with certain neurophysiological properties "it would be necessary to show that the introspective observations reported by the subject can be accounted for in terms of processes which are known to have occurred in his brain" (Place

1956/2002).[47] Although Place is not entirely clear, a natural way of understanding him is saying that if the neurophysiological properties of a subject causally explain these introspective reports, we have good grounds for accepting the identity of phenomenal properties with neurophysiological ones. Along with his stated purpose of showing that it is a meaningful scientific hypothesis that phenomenal properties are neurophysiological properties, he remarks, "We can identify consciousness with a given pattern of brain activity, *if we can explain the subject's introspective observations by reference to the brain processes*" (my emphasis) (Place 1956/2002, 55). Furthermore, there would be no need to suppose that these introspective reports are a result of special non-physical objects and events within a private "phenomenological field".[48] A reduction of phenomenal properties to neurophysiological properties is grounded partly on the claim that these properties have certain physical effects.

Smart's (1959) defense of reductionism begins by offering rejoinders to several objections to the identity of phenomenal and neurophysiological properties. This negative component of his overall defense is crucial, since the only positive reasons he offers in support of the reduction is an appeal to Occam's razor. Near of the end of his essay he remarks,

> If it be agreed that there are no cogent philosophical arguments which force us into accepting dualism, and if the brain process

[47] It is worth being clear that the focus in most early discussions of reductionism is on the identity of experiential mental properties (i.e., phenomenal properties) and neurophysiological properties. It seems that many of the early advocates of the identity theory were convinced of a behaviorist analysis of intentional properties but not for phenomenal ones (Smart 2002/1956, 55). These phenomenal properties, then, had to be accounted for in a different way.

[48] See (Place 1956/2002, 59) for the "phenomenological fallacy".

> theory and dualism are equally consistent with *the facts*, then the principles of parsimony and simplicity seem to me to decide overwhelmingly in favor of the brain-process theory (my emphasis) (Smart 1959/2002, 67).

Although Smart does little to give us an idea of what sorts of facts are amongst "the facts", we can safely assume various kinds of causal facts would be included. We might, then, be able to extract the following kind of argument from Smart's discussion. Experiences of yellowish orange after-images typically have physical causes $\{C_1, C_2, \ldots, C_n\}$ and typically have physical effects $\{E_1, E_2, \ldots, E_n\}$. Both the reductionist and the dualist explain equally as well these typical causes and effects of yellowish orange after-images. Given Occam's razor, there is reason to endorse the reductive theory over the dualist theory. Once again, we can see how a reduction of mental properties to neurophysiological properties is grounded partly on claims that mental properties are causal properties with certain physical effects.

Both Place's and Smart's arguments in favor of reductionism proceed relatively free from *a priori* considerations concerning our mental concepts. Placing them within a more contemporary dialectic, we might identify them as endorsing the kind of *a posteriori* physicalism advanced by Ned Block and Robert Stalnaker (1999).[49] This is the feature that places them at odds with the second tradition of arguing for reductionism exemplified by David Lewis (1966, 1980), David Armstrong (1981) and more recently elaborated and defended by Kim (1993a, 2005).

[49] Another way of putting this is that Place and Smart might be thought to endorse the view that mental/neurophysiological identities are in no need of explanation while reductionists like Lewis, Armstrong, and Kim think mental/neurophysiological identities can be further explained.

Lewis (1966) begins with the premise that the definitive characteristic of any experience is its causal role (Lewis 1966, 19). This amounts to saying that our concept of pain is defined in terms of its typical causes and its typical effects.[50] Lewis (1980) elaborates on these points by claiming that the causal role definitive of phenomenal properties (and any mental property for that matter) is the role attributed to it by our *folk psychology*. The second premise of his argument is interestingly similar to Completeness. It is the plausible scientific hypothesis that,

> There is some unified body of scientific theories, of the sort we now accept, which together provide a *true and exhaustive account of all physical phenomena* (i.e., all phenomena describable in physical terms). They are unified in that they are cumulative: the theory governing any physical phenomena is explained by theories governing phenomena out of which that phenomenon is composed and by the way it is composed out of them. The same is true of the latter phenomena, and so on down to fundamental particles of fields governed by a few simple laws, more or less as conceived of in present day theoretical physics (my emphasis) (Lewis 1966, 23).

Lewis's second premise involves more than just the claim that every physical phenomenon that has a cause has a sufficient physical cause. It also incorporates the claim that all physical phenomena have a "vertical" explanation in terms of lower-level physical phenomena.[51] The result is that it is quite likely some physical property or other will occupy the causal role definitive of each mental property.

---

[50] Lewis goes further and claims that "by analytic necessity these conditions are true of the experience and jointly distinctive of it" (Lewis 1966, 20). Although, Armstrong (1981) agrees that our concept of a mental property is a concept of a property with certain sorts of causes and effects, he does not go as far as Lewis. Despite his claim that this is a conceptual analysis of our concept of a mental property, he explicitly rejects that such analyses are meaning-giving: "I think that sentence translation (with synonymy) is too strict a demand to make upon purported conceptual analysis. What more relaxed demand can we make and still have a conceptual analysis? I do not know" (Armstrong 1981/2002, 85).

[51] A "vertical" explanation is suppose to be a reductive explanation without any presumption that such an explanation involves an *a priori* element. These "vertical" explanations are contrasted to

The argument, then, is rather straightforward: if it is an analytic truth that mental property M occupies causal role R and some physical property P occupies causal role R, then M is identical to P. The identification of mental properties with some physical property is not grounded in theoretical considerations like Occam's razor, but instead follows from the semantics of theoretical terms and the plausible scientific hypothesis quoted above (Lewis 1980/2002, 88). Additionally, Lewis thinks there is little reason to doubt that the physical properties that occupy these roles are the properties cited in neurophysiological theory, since "we have no notion of any other physical phenomena that could possibly occupy them, consistent with what we know" (Lewis 1966, 24). Not only are mental properties physical properties, but, the argument concludes, they are *neurophysiological* properties.

Both traditions of arguing for reductionism place claims of mental causation at the forefront. It is because mental properties are taken to have these kinds of causes and these kinds of effects that we should accept their identity with neurophysiological properties. However the reductionist prefers to defend the identity of mental properties with neurophysiological properties, the dissolution of the exclusion problem is the same – there is no competition without distinct causal competitors.

This is not an essay on the faults, problems, heresies, or sacrilege of reductionism. In fact, I have strong sympathies with the kinds of causal arguments

---

"horizontal", causal explanations. See Ch. 4 and 5 of (Kim 2005) for a great discussion of "vertical" explanations and an argument that all such explanations must be reductive (i.e., involve an *a priori* element).

discussed above for a reduction of mental to neurophysiological phenomena. I have always felt that the best reasons in favor of a reduction of mind are considerations having to do with causation. The present essay, however, is an approach to the exclusion problem from a *nonreductionist* perspective. This is, of course, not an undisputed perspective to take, but it is nonetheless the perspective I have elected to explore these issues from. The motivation for this has mainly to do with the potential for a *dialectical shift* in the philosophy of mind. If it can be shown that the exclusion problem can be avoided while accepting Nonreductionism, then what we have are good reasons for thinking that the causal argument is mistaken. The impetus for reductionism would thereby be frustrated. A secondary goal of this essay is to show that reductionism is frustrated in precisely this way.

Chapter 3

Proportionality and the Exclusion Problem

Late for a meeting, Suzy decides the quickest way to get to her office is to hail a cab. This decision comes after a brief episode of deliberation, since several options were available for Suzy to take. Commonsense tells us that Suzy's deliberation causally led to her decision to hail a cab which is causally responsible for her cab hailing behavior. But, given the principle of Completeness, her cab hailing behavior has a sufficient neurophysiological cause. However, the nonreductive physicalist must admit that her decision to hail a cab is distinct from the neurophysiological cause of her cab hailing behavior. Therefore, by Exclusion, this mental phenomenon is either causally excluded by some neurophysiological phenomenon or limited to being a mere causal overdeterminant. The nonreductive physicalist, therefore, faces a serious dilemma: epiphenomenalism or overdetermination. The principal challenge for the nonreductive physicalist is how to avoid *superfluous* mental causation without diving headlong into epiphenomenalism. How is it possible for anyone besides the reductionist to eschew both horns of the exclusion dilemma?

This chapter serves as a bridge between the typical presentation of the exclusion problem and the solutions I find most promising, which employ importantly different causal notions. In Section 3.1, I introduce the distinction between the notions of *sufficiency* and *relevance* as opposed to *difference-making*. Section 3.2 reformulates the exclusion problem in terms of difference-making and

then explores Stephen Yablo's and Sydney Shoemaker's solution to the problem. The idea unifying these authors is that genuine causation is not sufficiency or relevance, but difference-making. For Yablo and Shoemaker, once we take the difference-making idea seriously, we will discover a rather straightforward solution to the exclusion problem. Specifically, each author claims that the reformulated version of Completeness is false: some bodily effects do *not* have difference-making neurophysiological causes. Their rejection of this amended version of Completeness thwarts the exclusion dilemma before it can even get started. In Section 3.3, I criticize Yablo and Shoemaker's solution to the exclusion problem which motivates a search for an alternative solution that takes the difference-making idea seriously.

Section 3.1: Causation as Difference-Making

I have followed the philosophical precedent and framed the exclusion problem in terms of the question of how the mental can be *causally sufficient for* or *relevant to* some bodily effect if the neurophysiological is already sufficient for and relevant to those effects. The notions of causal sufficiency and relevance are so pervasive in the literature on the exclusion problem that most formulations of the assumptions that generate the dilemma are strictly in those terms. For instance, although formulations of Exclusion vary slightly from author to author, the central idea is relatively clear: the *causal sufficiency* of x for an effect y threatens the causal status of anything x* appropriately related to and distinct from x *vis-à-vis* y. Roughly, the threat is either that the *causal relevance* of x* for y is effectively

removed by x or x limits x*'s role to that of a causal overdeterminant (Yablo 1992b, fn. 53, 273).

However, as the literature on causation and explanation has made clear, we should be unsatisfied with vindicating the mere causal sufficiency or relevance of mental phenomena.[52] A famous example given by Wesley Salmon (1984) nicely illustrates that a condition can be sufficient for an effect, but fail to be a cause of that effect. Suppose that Jones is a man who takes birth control pills and, therefore, fails to become pregnant. Not only is it true that all men who take birth control pills fail to become pregnant, but this is *non-accidentally* true. There is a lawful connection between men who take birth control pills and the failure to become pregnant. But it should be clear that Jones's taking of birth control pills is not a cause of his failure to become pregnant. As we are likely to say, Jones would have failed to get pregnant *regardless* of whether he took birth control pills. His taking of the pills made no difference to his failure to become pregnant, but rather it was his lack of the requisite biology. Furthermore, many discussions of the exclusion problem make heavy weather of the distinction between causal relevance and causation. For example, Yablo remarks that "even if some mental antecedent is causally relevant, it is a further question yet whether it actually *causes* the effect" (Yablo 1992b, 273).[53] Peter Menzies (2008) echoes this distinction between causation and causal relevance when he says, "Causal relevance … is a loose and

[52] Here we are to understand causal sufficiency as one event determining the occurrence of another or, if indeterminism is true, fixing the objective probability of the occurrence of another (Yablo 1992b, fn. 7, 247). Given the laws of nature, the occurrence of this condition fixes the objective probability of or determines the occurrence of this other condition.

[53] I'll discuss the details of Yablo's account of causation further in Section 3.2.

undiscriminating concept … By contrast, the concept of causation is more discriminating" (Menzies 2008, 200).

The reason we should be unsatisfied with just the causal sufficiency or relevance of the mental is that neither guarantees that the mental *makes a difference*. And finding a way for the mental to make a difference in a world that is fundamentally physical is the real prize for the nonreductive physicalist. Consider Norman Malcolm's (1968) early presentation of the exclusion problem,

> The movements of the man on the ladder would be completely accounted for in terms of electrical, chemical, and mechanical processes in his body … Given the antecedent neurological states of his bodily system together with general laws correlating those states with the contractions of muscles and movements of limbs, he would have moved as he did regardless of his desire or intention (Malcolm 1968, 53).

Malcolm's concern is that the thesis of "mechanism" threatens the difference-making causal status of the mental. If the man had moved as he did *regardless of his mental condition*, then the mental made no difference to the man's actions. A true vindication of mental causation is finding a way for the mental to be *a cause*, something "that 'makes the difference' between the effect's occurring and its not" (Yablo 1992b, 274) (see also (Menzies 2008, 205)).

Section 3.2: Intensive Parts and Proportional Causes

Let us proceed forward with Yablo and consider the exclusion problem from the perspective in which *real* mental causation is having the mental *make a difference*. In lieu of this, the primary worry for the nonreductive physicalist is finding a way for the mental to make a difference in a physical world that is

causally complete. If every bodily effect that has a cause has a difference-making neurophysiological cause, the mental appears forced into being a second and superfluous difference-making cause of its effects. How does mental causation not introduce overdetermination? Yablo avoids the threat of redundant mental causation by denying an appropriately amended version of Completeness, that is, it follows from his view that not every bodily effect has a difference-making neurophysiological cause. The same general solution to the exclusion problem follows from Sydney Shoemaker's (2003c, 2007) recent account of realization. In this section, I discuss Yablo and Shoemaker's view that mental phenomena are *intensive parts* of their neurophysiological realizers which eventually leads them to accept that causes must be *proportional* to their effects. The idea of proportionality is proposed as an analysis of difference-making which, when applied to cases of mental causation, generate their solution to the exclusion problem.

A few preliminary remarks are in order before moving forward. First, both Yablo and Shoemaker take it for granted that events are *fine-grained*. I shall understand a commitment to fine-grained events as a thesis about the denotations of event nominals. Specifically, events are fine-grained if a nominal's denotation changes when (a) it's gerund is modified by an adjective (e.g., 'Suzy's **throwing** of the rock'/'Suzy's **abrupt throwing** of the rock'), (b) it involves a more or less specific gerund (e.g., 'Socrates's **drinking** of the hemlock'/'Socrates's **guzzling** of the hemlock'), or (c) it involves focus (e.g., 'Socrates's *drinking of the hemlock* at noon'/'Socrates's drinking of the hemlock *at noon*'). Yablo and Shoemaker at least

take the nominal pairs in (a) and (b) to have different denotations so I shall say that they hold a fine-grained view of events.

Second, as will become clear in Section 3.2.1, Yablo explicitly argues for the conclusion that the mind-body relation is a species of the determinable-determinate relation. Shoemaker, on the other hand, talks of the mind-body relation as realization, but often notes the close resemblance between realization and the determinable-determinate relation. Consequently, Yablo and Shoemaker take mental event nominals as involving gerunds less specific than neurophysiological event nominals. For instance, the nominal 'Socrates's pain' is less specific than 'the firing of Socrates's C-fibers' just as 'Socrates's **drinking** of the hemlock' is less specific than 'Socrates's **guzzling** of the hemlock'. Given their fine-grained view of events, it follows that mental event nominals do not co-refer with neurophysiological event nominals.

Finally, Yablo and Shoemaker often talk of property-instances entering into causal relations. I interpret them as holding the metaphysical thesis that events are identified with property-instances where these are exemplifications *of* properties *by* objects *at* times.[54] On this view, events are structured particulars where a difference, for example, in the constitutive property entails a difference in the event. Socrates's drinking of the hemlock is a distinct event from Socrates's guzzling of the hemlock, since the constitutive properties of drinking and guzzling are distinct.

[54] See (Kim 1976) and (Goldman 1970).

This view of events makes them very nearly as fine-grained as facts (i.e., true propositions).

Section 3.2.1: Determination and Realization

As I understand them, both Yablo and Shoemaker maintain that mental properties and their instances are intensive parts of their neurophysiological realizers. The notion of intensive parthood is purely modal and can be defined as follows: x is an *intensive part* of y just in case y entails x but x does not entail y (McLaughlin 2007, 159). Let us call the thesis that mental properties and their instances are intensive parts of their neurophysiological realizers the *intensive parthood thesis*. Yablo and Shoemaker argue for the intensive parthood thesis in different ways. It is a direct consequence of Yablo's view that the mind-body relation is a species of the determinable-determinate relation; while, Shoemaker argues that it follows from his most recent account of realization.

The primary reason we should construe the mind-body relation as a species of the determinable-determinate relation is that it explains the "reigning orthodoxy" in the philosophy of mind that "the mental is *supervenient* on, but *multiply realizable* in, the physical" (Yablo 1992b, 254). Mental properties are said to be supervenient on neurophysiological properties such that necessarily, for every x and every mental property M of x, x has some neurophysiological property N such that necessarily all N's are M's. In addition to being supervenient properties, mental properties are also said to be multiply realized properties in which necessarily, for every mental property M, and every neurophysiological property N

which necessitates M, possibly something is an instance of M without being an instance of N. Together these claims paint a picture of mental properties as *asymmetrically necessitated* by their neurophysiological realizers (Yablo 1992b, 256). The neurophysiological guarantees the mental, but the reverse is not the case.

As Yablo makes clear, this kind of asymmetric necessitation is guaranteed to hold if we construe mental properties as determinables of their neurophysiological realizers. An instance of some determinate property is an instance of its determinable property, not *simpliciter*, but in a specific way (Yablo 1992b, 252). Consider, for example, the determinable property red and its determinate scarlet. Every instance of scarlet is a specific way of being red such that (a) necessarily, if something is scarlet then it is red and (b) it is possible for something to be red without being scarlet (say, by being crimson). This relation of determination guarantees that determinables supervene on and are multiply realized in their determinates. Adopting the hypothesis that mental properties are determinables of their neurophysiological realizers enables a straightforward explanation for why the mental is both supervenient on and multiply realized in the neurophysiological. But, more importantly, we are assured that mental properties are intensive parts of their neurophysiological realizers. Determinates entail their determinables, but the reverse is not the case.

Shoemaker (2001) arrives at the intensive parthood thesis by adopting a new definition of realization. The traditional account of realization is associated with the functionalist idea of *role occupancy*. Realized properties are second-order

properties and their realizers first-order. A property P is said to realize a distinct property Q just in case P is the first-order property that occupies the causal role assigned to the second-order property Q by either folk or scientific psychology. Shoemaker however resists this account of realization, attributing to it a variety of problems, in particular, that it leads to either overdetermination or epiphenomenalism for mental properties (Shoemaker 2001, 76 – 77). As a replacement, Shoemaker (2001) stipulates, with some minor modifications given in (2007), that P *realizes* a property Q just in case the set of forward-looking conditional powers of Q is a proper subset of the set of forward-looking conditional powers of P.[55]

Shoemaker explains the notion of a conditional power,

> Any property whose instantiation can be a cause or partial cause of something will be such that its instantiation bestows on its subject a set of what I call 'conditional powers'. A thing's having a power simpliciter is a matter of its being such that in certain circumstances, for example, its being related in certain ways to other things of certain sorts, causes (or contributes to causing) certain effects. A thing has a conditional power if it is such that if it had certain properties it would have a certain power simpliciter, where those properties are not themselves sufficient to bestow that power simpliciter (Shoemaker 2001, 77).

[55] The modifications first offered in (Shoemaker 2007) won't matter much for our discussion but include that (a) the set of backward-looking conditional powers of P is a proper subset of the set of backward-looking conditional powers of Q and (b) that P is not a conjunctive property having Q as a conjunct. The forward-looking conditional powers of P are powers of P such that if P were instantiated in some object *along with certain other properties* then the instance of P would cause or contribute to causing such and such effect. The backward-looking conditional powers of P are powers of P such that if *certain other properties were instantiated* then an instance of P would be an effect. Roughly, forward-looking conditional powers of P are features of P having to do with "how the instantiation of the property contributes to producing various sorts of effects" and backward-looking conditional powers of P are features of P having to do with "what sorts of states of affairs can cause the instantiation of the property" (Shoemaker 2007, 12).

For example, the property of *being knife-shaped* bestows on objects that possess it the conditional power of being capable of cutting butter *if it is made of a certain material* (e.g., wood, iron, steel, etc.). Therefore, if an object possessed both this property and the property of *being made of steel*, it would have the power simpliciter to cut butter.

When one property P realizes another property Q, P and Q literally share their conditional powers. The conditional powers of Q are a *proper subset* of the conditional powers of P. The realizer property will, more than likely, have conditional powers that go beyond those of the realized property, but any conditional power of the realized property is also a conditional power of the property that realizes it. Take, for instance, the properties red and scarlet which Shoemaker considers to be a paradigmatic case of one property realizing another (Shoemaker 2001, 78). Sophie the pigeon is trained to peck at red things to the exclusion of non-red things. The property of being red confers upon objects that possess it the conditional power of "evoking a pecking response in the likes of Sophie" (Shoemaker 2001, 78). And since the determinate property scarlet realizes the property redness, it is also true that scarlet bestows upon objects that possess it this very same conditional power.[56] But now imagine that Sophie has a sister named Alice who has been trained to peck at scarlet things and not at other shades of red. The property of scarlet bestows upon objects that possess it the conditional

[56] For Shoemaker, this does not entail that the scarlet property-instance *causes* Sophie's pecking behavior because causes must be *proportional* to their effects and scarlet fails to satisfy this condition. See Section 3.2.2 for a discussion of this constraint on causation and the role it plays in providing a solution to the exclusion problem.

power of evoking a pecking response in the likes of Alice, and this is a conditional power that the property redness lacks. An instance of the property scarlet realizes an instance of red because the conditional powers of redness are a proper subset of the conditional powers of scarlet.

Elaborating further Shoemaker claims that,

> The instantiation of the determinate entails the instantiation of the determinable, and can be quite naturally said to include it. It seems natural to say that being scarlet is in part being red. Likewise, the instantiation of a realizer property entails, and might naturally be said to include as a part, the instantiation of the functional property realized (Shoemaker 2001, 81).

This suggests that Shoemaker sees the relationship between mental properties and their realizers as similar to that between determinables and their determinates. Realizer properties *include* the properties they realize as a *part* just as determinates include their determinables as parts. Shoemaker claims that his alternative account of realization guarantees that realized properties are intensive parts of their realizers: if P realizes Q, then an instance of P *entails* an instance Q, but not *vice versa* (Shoemaker 2001, 94 – 95).[57]

Section 3.2.2: The Proportionality Thesis

Suppose the nonreductive physicalist accepts Yablo's and Shoemaker's thesis that mental properties and their instances are intensive parts of their neurophysiological realizers. Exactly how does this help them avoid the threat of overdetermination? After all, an instance of a mental property is distinct from the instance of its neurophysiological realizer regardless of the fact that the former is

[57] However, see (McLaughlin 2007, 159 – 161) for arguments that Shoemaker's account of realization does not guarantee that realized properties are intensive parts of their realizers.

an intensive part of the latter. When a mental property instance causes some bodily effect, what prevents its intensive whole – the neurophysiological property-instance that realizes it – from also counting as a cause? What rules out that we once again have redundant causation of the bodily effect?

In order to answer this question, both Yablo and Shoemaker claim that causes make a difference to their effects and that this should be understood as causes being *proportional* to their effects.[58] The idea here is that causes "should incorporate a good deal of causally important material but not too much that is causally unimportant" (Yablo 1992b, 274). That is, "causes should be specific enough but no more specific than is required to make the difference to their effect" (Menzies 2008, 209). A part is sometimes better suited to being a cause than a whole *because* it is the part rather than the whole that is proportional to the effect. Let us call the thesis that causes must be proportional to their effects the *proportionality thesis*. Although Yablo defines proportionality in terms of four distinct notions, it will only be necessary to discuss two: the proportionality thesis entails that c causes e only if (a) c is *required* for e and (b) c is *enough* for e.[59]

It is useful to employ a terminology introduced by Matthew McGrath (1998) to capture the notions of required and enough. Let us define *screening off* as follows:

---

[58] See (Yablo 1992b, 274), (Shoemaker 2001, 93), and (McLaughlin 2007, 168).

[59] There are two other conditions Yablo puts on causation that will not be necessary for our discussion. The first he calls the *contingency condition* and states that effects must be counterfactually dependent on their causes in the sense of Lewis (1973a). See Section 3.4.1. The second he calls the *adequacy condition* which states that if the cause had not occurred then if it had, the effect would have occurred as well.

> (Screening Off) *x screens off y* from an effect e just in case if *x* had occurred without *y* then e would still have occurred.

Now Yablo defines the required condition as follows[60]:

> (Required) *x* is *required for* an effect e just in case none of *x*'s determinables screens off *x* from e.

If we prefer Shoemaker's terminology we can say that x is required for an effect e just in case nothing *realized by* x screens it off from e. We can illustrate this condition by considering a scenario in which Socrates guzzles some hemlock and dies. Furthermore, we stipulate that Socrates could not have drunk the hemlock without guzzling it. He was, to the dismay of Xanthippe, a notoriously sloppy drinker. Yablo remarks,

> Intuitively, it appears that not *all* of the guzzling was needed, because there occurred a lesser event, the drinking, which would still have done the job even in the guzzling's absence. By hypothesis, of course, without the guzzling this lesser event would not have taken place; but that doesn't stop us from asking what would have happened if it had, and evaluating the guzzling on that basis (Yablo 1992b, 276).

In the described scenario, Socrates's guzzling of the hemlock is not required for the death, since there is a determinable of the guzzling – the drinking – that screens it off from the death. The drinking screens off the guzzling *because* had the drinking occurred in the absence of the guzzling, the death would still have occurred. An event c is proportional to an effect e only if c is required for e, that is, none of c's determinables screens it off from e.

Next, Yablo defines the enough condition as follows:

---

[60] See (Yablo 1992b, 276) for the original terminology.

> (Enough) x is *enough for* an effect e just in case x screens off all of its determinates from e.

Once again, if we prefer Shoemaker's terminology we can say that x is enough for an effect e just in case x screens off all of its *realizers* from e. To illustrate this condition, suppose that a safety valve is connected by a pipe to a boiler. The valve mechanism stiffens due to a preexisting structural defect. This stiffening of the safety mechanism slows down the opening of the valve enough so that the pressure builds up in the boiler and it explodes. The opening *per se* of the valve is not a cause of the explosion, since intuitively "the effect required something more" (Yablo 1992b, 277). In this case, the opening *per se* is not enough for the explosion of the boiler, since there is a determinate of the opening *per se* – the slow opening – that is not screened off from the explosion. The opening *per se* does not screen off the slow opening *because* had the opening *per se* occurred without the slow opening, the explosion would not have occurred. An event c is proportional to an effect e only if c is enough for e, that is, c screens off all of its determinates from e.

Section 3.2.3: A Solution to the Exclusion Problem

Yablo and Shoemaker are each committed to the intensive parthood thesis and claim that sometimes parts rather than wholes are better suited for being a cause of some effect. This depends on their views about causation, in particular, that causes make a difference and so must be proportional to their effects. If the proportionality thesis is correct, there are some interesting consequences for the exclusion problem.

Let us first consider a case in which we *assume* that some bodily effect b has a difference-making mental cause m and that m is realized by some neural activity n. Let us say that Suzy arrives on Jones's doorstep and decides to ring the doorbell. Under the conditions in which Suzy's decision is a cause and so is proportional to the ringing behavior, it follows that the neural activity that realizes this decision is *excluded* from being a cause of ringing behavior. The proof of this is as follows:

(1) x is *enough* for y iff x screens off all its determinates from y (definition of Enough),
(2) x is *required* for y iff none of x's determinables screens it off from y (definition of Required),
(3) m is enough for movement b (assumption),
(4) therefore, m screens off all its determinates from b (by 1, 3),
(5) n is a determinate of m (Yablo's mind-body relation),
(6) therefore, m screens off n from b (by 4, 5),
(7) therefore, n is not required for b (by 2, 6).

If causes must be proportional to their effects, the fact that Suzy's decision is a cause of her ringing behavior entails that the neurophysiological realizer of this decision fails to be proportionate to her ringing behavior.[61] Generally, if some mental property-instance is *enough* for an effect, then its realizer cannot be *required* for it (Yablo 1992b, 278). If we assume that the intensive part *makes the causal difference*, then given the proportionality thesis it follows that the intensive whole does not. This vindicates an exclusion principle Menzies (2008) calls "downward" exclusion:

[61] If we define realization as Shoemaker does, then the fact that the neurophysiological realizer of Suzy's decision fails to count as a cause of her ringing behavior does not entail that this realizer lacks the conditional power to bring about the ringing behavior. In other words, the intensive whole that is the realizer of Suzy's decision still has all the same conditional powers as Suzy's decision, but the whole fails to be proportionate to Suzy's ringing behavior.

> (Downward Exclusion) If some mental property-instance m is a difference-making cause of e, then no neurophysiological property-instance n that realizes m is a difference-making cause of e.

The truth of "downward" exclusion is incompatible with m and n causally overdetermining the effect e, since it is false that both m and n are causes of e. If causes must be proportional to their effects, then mental *causes* guarantee that their realizers are not causes. The worry that mental causation always introduces a superfluous cause into the story is entirely misguided once causation is understood to be a proportional relation.

We arrive at a similar result when we consider a case where it is *assumed* that some bodily effect has a difference-making neurophysiological cause. This time it follows that the mental property-instance fails to be a cause. The proof is similar to the one above:

(1) x is *enough* for y iff x screens off all its determinates from y (definition of Enough),
(2) x is *required* for y iff none of x's determinables screens it off from y (definition of Required),
(3) n is required for b (assumption),
(4) therefore, none of n's determinables screens it off from b (by 2, 3),
(5) m is a determinable of n (Yablo's mind-body relation),
(6) therefore, m does not screen off n from b (by 4, 5),
(7) therefore, m is not enough for b (by 1, 6).

If the neurophysiological realizer of a mental property-instance is required for a bodily effect, the mental property-instance cannot be enough for it. This conclusion vindicates another exclusion principle Menzies (2008) has called "upward" exclusion:

> (Upward Exclusion) If some neurophysiological property-instance n is a difference-making cause of e, then no mental property-instance m realized by n is a difference-making cause of e.

Once again, the proportionality thesis has ruled out the possibility of overdetermination.

But what about a case in which both the mental and neurophysiological property-instances purport to be a cause of some bodily effect? It would be wrong to interpret Yablo as holding that the mental property-instance *always* counts as a cause under these circumstances. Instead which property-instance is really a cause of the effect depends on which satisfies the proportionality constraints placed on causation. As it turns out, this is something that depends upon the finer empirical details of the case. Generally, there is mental causation whenever the occurrence of the bodily effect is "relatively insensitive to the finer details" of the mental's neurophysiological realization (Yablo 1992b, 278). And often the neurophysiological realizers of mental property-instances are "overladen with materials to which the effect is in no way beholden" (Yablo 1992b, 279). This is another way of saying that sometimes an intensive whole is laden with causal details that are superfluous to the effect. In such cases, it will be some intensive part of the whole that is proportionate to the effect. But, again, determining that this is true for any particular bodily effect requires careful evaluations of the counterfactuals associated with the *required* and *enough* conditions. Regardless of which property-instance ends up being a cause, we do know that only one can –

overdetermination of an effect by a mental property-instance and an instance of its realizer is impossible given the proportionality thesis.

Before moving on let me deflect the worry that mental property-instances are never *enough* for intentional bodily movements. Suppose that Suzy's decision occurs, but is realized in some radically different way, say, in ectoplasm. It is not obvious that had the decision been realized in this way that the ringing behavior *would* still have occurred. Perhaps in such a case the behavior *might* not have occurred. Yablo reminds us that worlds in which the decision is realized in some radically different way are not worlds relevant to evaluating the *enough* counterfactual. All that matters are "the *nearest* world[s] where its physical implementation was not as actually – the world[s] in which it undergoes only the minimum physical distortion required to put its actual implementation out of existence" (Yablo 1992b, 278). The question of whether the ringing behavior might not have occurred if the decision had been realized in ectoplasm does not need to be answered. All we need to answer is whether the ringing behavior would still have occurred in the *nearest worlds* where it is realized differently, that is, if the decision had been realized in a slightly different neurophysiological way. And this looks to be a more tractable question and one we can, for the most part, be confident in answering.

In conclusion, Yablo's and Shoemaker's solution to the exclusion problem is fairly simple: if there is a mental cause of some bodily effect, then its neurophysiological realizer *cannot* be a cause of that effect and so

overdetermination is avoided. A proper understanding of causation, namely that it requires causes to be proportional to their effects, ensures that *mental causation* does not introduce overdetermination into the causal story. However, Yablo and Shoemaker must deny that every bodily effect that has a cause has a neurophysiological difference-making cause. If some bodily effect has a mental cause, then from the principle of "downward exclusion" it follows that the realizer of this mental cause cannot itself be a cause.

Nonetheless, rejecting this version of Completeness does not threaten the claim that every bodily effect has a *sufficient neurophysiological cause*, since everything that happens is still a strict "causal consequence of its physical antecedents" (Yablo 1992b, 279). That is, mental causation without overdetermination is consistent with there being some neurophysiological condition of the brain and nervous system that determines or fixes the objective probability of the occurrence of the bodily effect. Once causal sufficiency is distinguished from causation, we can see that our mistake and the primary source of confusion in discussions of the exclusion problem is to assume that the "outcomes of the kind normally credited to human agency are *caused* by their physical antecedents" (my emphasis) (Yablo 1992b, 280). The effects of mental causes are not overdetermined because they fail to have neurophysiological *causes*.[62]

[62] Peter Menzies (2008) also endorses this response to the exclusion problem. In the final paragraphs of his paper he remarks, "Acceptance of the new version of the exclusion principle does not automatically compel us to accept the conclusion of the exclusion argument to the effect that mental properties do not cause physical properties. However, the plausibility of the new exclusion principle does mean that the critical spotlight needs to be shifted to the other crucial principle of a reformulated version of [the] argument – the strengthened causal closure principle" (Menzies 2008,

Section 3.3: Problems for the Proportionality Thesis

The notions of realization, causal relevance, and causal sufficiency are importantly different from the notion of causation. The former are technical philosophical terms and authors are allowed to stipulate how they understand them without having to test them against our considered judgments. The only reason we have to adopt stipulated definitions of technical terms depends on their potential to resolve a range of philosophical problems. But the notion of causation is not like this. There are fairly robust pre-reflective intuitions we have about specific cases and these judgments, for the most part, ought to be respected. There is, of course, no implication that these intuitions are sacrosanct and therefore unrevisable. My point is only to draw attention to the fact that our theory of causation should incorporate our considered judgments about causation as much as possible. In this section, I argue that if the proportionality constraint is consistently applied, it leaves few pre-reflective causal judgments intact and, therefore, ought to be rejected. Consequently, a rejection of proportionality on these grounds undermines its utility in offering a solution to avoid the threat of causal overdetermination.

The proportionality thesis states that it is part of the truth conditions for causal claims that causes are *proportional* to their effects. Specifically, Yablo and Shoemaker propose that 'c causes e' is true only if c is *required* for e and c is *enough* for e. The idea is that causes make a difference to their effects because

---

216). Menzies argues that on his account of causation, just as on Yablo's and Shoemaker's, it is reasonable to maintain that intentional bodily movements will sometimes have mental causes but *no neurophysiological causes*.

they are specific enough, but not too specific to bring about their effects. We can state their partial truth conditions as follows:

> (PC) 'c causes e' is true only if (a) no determinable of c screens it off from e and (b) c screens off its determinates from e.

Making the underlying counterfactual claims explicit, (PC) says that the following counterfactuals must hold in order for 'c causes e' to be true: (a) for every determinable d_1 of c, if d_1 had occurred without c, then e would not have occurred; and (b) for every determinate d_2 of c, if c had occurred without d_2, then e would still have occurred. These counterfactuals are at the heart of the proportionality thesis.

Let us first consider a pair of causal claims familiar from Yablo's exposition of proportionality in which Socrates guzzles some hemlock and dies:

(1a) Socrates's drinking of the hemlock caused his death.

(1b) Socrates's guzzling of the hemlock caused his death.

Yablo considers (1a) and (1b) together and stipulates that the guzzling "contributed nothing" to Socrates's death, so that (1a) is true while (1b) is false (Yablo 1992b, 275). Yablo's truth conditions for causal claims demand that causes be *required* for their effects and so the truth of (1a) entails that no determinable of the drinking screens it off from the death. This, however, is not the case. Consider the claim that:

(1c) Socrates's ingesting of the hemlock caused his death.

The ingesting of the hemlock is a determinable of the drinking that screens it off from the death *because* if Socrates had ingested the hemlock without drinking it, he

would still have died. We are strongly inclined to judge (1a) true and (PC) does not respect this intuition. This provides some *prima facie* reason for rejecting the view that causes must always be required for their effects.

The proportionality thesis captures the idea that "causes should be specific enough but not more specific than is required to make the difference to their effect" (Menzies 2008, 209). In addition to threatening the truth of causal claims like (1a) and (1b), this idea also threatens the truth of (1c) for the very same reasons. The ingesting of the hemlock is, contrary to what we intuitively think, not a cause of the death. There is a determinable of the ingesting – namely, doing something fatal – that screens the ingesting off from the death (Bontly 2005, 340). If Socrates had done something fatal without ingesting the hemlock, he would still have died.

The issue here is more problematic than it might first appear. (PC) commits one to holding, not just that token-causal claims like (1a) – (1c) are false, but that *all* token-causal claims of this general sort are false. Suppose that Plato drinks some hemlock and dies. Given that every instance of drinking hemlock is invariably accompanied by an instance of its determinable doing something fatal, every instance of drinking hemlock is screened off from the death it purportedly causes. The result is that no instance of drinking hemlock will ever or has ever caused someone's death. It would be unreasonable to hold in spite of this that 'drinking hemlock tends to cause death' is a true causal generalization. So (PC) not only threatens token-causal claims like (1a) – (1c) but their associated causal generalizations as well. Drinkings, guzzlings, and ingestings of poisons do not tend

to cause deaths, since, given (PC), none of these are required for their purported effects. This is a significant cost to those who maintain that causes must be required for their effects and more *prima facie* reason to reject the proportionality thesis.

As I see it, the proportionality thesis forces us to reject many causal claims we think capture paradigmatic cases of causation. Suzy throws a rock at an empty glass bottle and it shatters. We are not hesitant in accepting (2):

(2) Suzy's throwing of the rock causes the shattering of the bottle.

But Yablo and Shoemaker must insist that, contrary to intuition, (2) is false, since there is some less specific event that is required for the effect. Perhaps *someone's* throwing of the rock, Suzy's throwing of *a hard object*, or even Suzy's *propelling* of the rock screens off Suzy's throw from the shattering. A related issue is that in most cases it will be practically impossible for us to identify a cause of an effect, since we will rarely be in a position to find an event that is *specific enough* but *no more specific than is required* to make a difference. This leads to a sort of *causal skepticism* about most, if not all, of our ordinary causal claims, since we cannot be sure we've found a cause until we can be sure we've found an event of the right specificity. I submit that a better alternative is to doubt the veracity of the proportionality thesis itself.

A related issue is that similar concerns arise if we accept only a *fine-grained view* of events and ignore the requirement of proportionality.[63] If Socrates's

[63] See (Dowe 2010) for a similar point.

drinking of hemlock is distinct from his guzzling of it which is distinct from his ingesting of it, and all of these events occur *at the same time* in the *same space* that Socrates occupies, then the question naturally arises which one of these events caused Socrates's death? If only one is a cause, it would be practically impossible for us to identify which one. There would have to be some constraint on causation by which to separate the non-causes from the cause and so we are lead back to something very much like the problematic proportionality thesis. Again, we would be forced to reject many, if not most, of our ordinary causal claims. If, on the other hand, all of these events are causes of Socrates's death, then it appears that the death is radically overdetermined, or involves a strange sort of "over"-causation. However, if we instead accepted that Socrates's drinking of the hemlock *just is* his guzzling it which *just is* his ingesting it, we wouldn't be forced into this dilemma. On a *coarse-grained view* of events, there is no need to appeal to the problematic proportionality constraint nor is there a worry about "over"-causation or causal skepticism.[64]

Proponents of proportional causation might respond that (PC)'s results indeed do not conform with many of our intuitive causal judgments and perhaps even leads to a sort of causal skepticism. But, the reply goes, none of this provides any reason to reject (PC). Instead, the defender of proportional causation may

[64] I think something like this argument provides solid grounds for preferring a coarse-grained view of events over a fine-grained view of events. To my mind, the crucial premise is that the death is radically overdetermined or involves a strange sort of "over"-causation. But why is this supposed to be an unwelcome consequence? If this argument works at all, there must be *some reason* why overdetermination is bad and ought to be avoided. This question assumes center stage in Chapter 5 as the claim that overdetermination is bad plays an important role in the exclusion argument. To foreshadow, I do not think *the kind* of overdetermination involved in this example is bad at all but rather is a pervasive feature of our world.

reasonably claim that their partial truth conditions are not intended as a "conceptual analysis" of our commonsense concept of causation. Conforming with intuition and avoiding causal skepticism are, at best, secondary concerns or, at worse, no concern at all. This reply certainly has some force, but it is not available to Yablo and Shoemaker. Their primary motivation for accepting the proportionality thesis is to solve the exclusion problem and this is only a problem if they are concerned to respect the "Moorean fact" of mental causation (Bontly 2005, 331). Mental causation is central to the commonsense picture we have of ourselves as agents and the exclusion problem is troublesome precisely because it threatens this picture. The fact that (PC) has such unintuitive consequences should be disconcerting for those like Yablo and Shoemaker who are inclined to defend the commonsense picture of mental causation. These problems are sufficient to motivate an alternative account of causes as difference-makers, one that conforms with our intuitive judgments better and offers new directions for avoiding the threats posed by the exclusion problem.

Chapter 4

The Exclusion Dilemma: The Epiphenomenalist Horn

I am a physicalist who maintains that there is mental causation and, specifically, the kind of mental causation that is able to sustain both *human agency* and *moral responsibility*. This implies that mental events are among the causes of our bodily movements. Furthermore, I accept the principle of Completeness, that is, I hold that in tracing the causal history of our bodily movements we will find, amongst this history, difference-making neurophysiological causes. But I am also a nonreductionist who holds that mental events are distinct from their neurophysiological realizers. Many have thought that the picture which emerges from these ideological commitments is a troublesome one. Take any piece of intentionally controlled behavior, some bodily movement we take to be an action. Amongst the causal history of this action, there will be a mental event *and that event's neurophysiological realizer*. But the principle of Exclusion tells us that if some neurophysiological event n causes an action b, then unless b is overdetermined no mental event m distinct from and existing simultaneously with n is such that it causes b. Therefore, I am forced into the *exclusion dilemma*: either I must reject mental causation and forsake both human agency and moral responsibility or embrace the view that mental events are, at best, redundant and superfluous causes.

In this chapter, I do *not* forsake mental causation. Instead, I offer a proposal for vindicating mental causation consistent with nonreductive physicalism. This

means that I am obligated to deal with the charge of overdetermination, which I tackle in Chapter 5. In Section 4.1, I reiterate David Lewis's simple theory of causation which takes counterfactual dependence among distinct events to capture the idea of difference-making. In Section 4.2, I endorse one component of Lewis's simple theory, namely that counterfactual dependence among distinct events is sufficient for causation, and claim that it provides us with the resources for grounding mental causation. My defense of this proposal, which I call *counterfactualism*, comes in two parts. The negative part fends off objections raised by Jaegwon Kim (2003, 2007) that counterfactual dependence cannot do the work required to ground mental causation. The positive part argues that, *at least compared to so-called productive accounts of causation*, the counterfactual account I offer is superior, since it is compatible with the empirical details of the physiological mechanisms of human action. Section 4.3 summarizes my arguments in this chapter and brings up several problems that remain for the counterfactualist to deal with. These problems are serious – some more so than others – but, regardless, a complete defense of counterfactualism must say something about them. Unfortunately, the present project must remain woefully incomplete.

Section 4.1: Lewis's Simple Theory

The fundamental idea of Lewis's (1973, 1979) simple theory of causation is that we can understand what it means for causes to make a difference to their effects in terms of counterfactual dependence,

> We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have

> happened without it. Had it been absent, its effects – some of them, at least, and usually all – would have been absent as well (Lewis 1973a, 557).

If we allow O(c) and O(e) to be the propositions that the events c and e occur respectively, then we can say that e counterfactually depends on c just in case the following counterfactuals are true:

(a) $O(c) \mathrel{\Box\!\!\rightarrow} O(e)$

(b) $\sim O(c) \mathrel{\Box\!\!\rightarrow} \sim O(e)$

The truth-conditions that Lewis (1973, 1979) proposes for counterfactuals involve a similarity relation amongst possible worlds. For any propositions A and C, the counterfactual $A \mathrel{\Box\!\!\rightarrow} C$ is true just in case either there are no possible A-worlds or every A-world amongst the set of most similar worlds is a C-world.

The (a)- and (b)-counterfactuals that define counterfactual dependence must be evaluated in a *non-backtracking* fashion.[65] Suppose Jim is driving well over the speed limit and comes upon an icy patch in the road. The speed and angle at which his car comes into contact with the ice causes him to momentarily lose control of the car threatening to send it into a ditch. However, Jim reacts swiftly and avoids the ditch. We think that if Jim had not reacted as swiftly as he did, the car would have ended up in the ditch. For instance, we think it is true that had Jim been drunk, he would not have regained control of the car in time and would have crashed into the ditch. But we might think the truth of this counterfactual is jeopardized by the following *backtracking argument*. We must not forget that Jim is a very

[65] I shall discuss this requirement more in Section 4.2.1.

responsible driver and would never have gotten behind the wheel had he been drunk. So if Jim had been drunk, he would not have crashed into the ditch, since he wouldn't have gotten into his car in the first place.

This reasoning involves *backtracking* and although Lewis thinks it is not always inappropriate, it is when we are trying to determine if one event is counterfactually dependent on another. The proper way to evaluate the (a)- and (b)-counterfactuals is by considering worlds where we suppose that A holds while *keeping fixed the past as much as possible*. In short, the most similar A-worlds are *not* those worlds where the past is substantially changed in order to make true A. The metric of weights and priorities Lewis eventually proposes for determining similarity amongst worlds in, e.g., *causal contexts* is intended to rule out backtracking evaluations.[66] Here I quote Lewis at length:

> (1) It is of the first importance to avoid big, widespread, diverse violations of law.
> (2) It is of the second importance to maximize the spatiotemporal region throughout which perfect match of particular fact prevails.
> (3) It is of the third importance to avoid even small, localized, simple violations of law.
> (4) It is of little or no importance to secure approximate similarity of particular fact, event in matters that greatly concern us (Lewis 1979, 472).

If the antecedent of the counterfactual is false, Lewis's metric directs us to consider whether the consequent holds in every world that has a history that pretty much exactly matches the history of the actual world except that a *small, localized*

[66] Hitchcock (2001) also points out that Lewis's metric requires that we "foretrack" when evaluating the relevant counterfactuals. This means that "if c causes e, we do not want to hold fixed e when evaluating the counterfactual 'If c had not occurred, then …'. If we do, the consequent of the conditional will obviously not be 'e would not have occurred'" (Hitchcock 2001, 275). So Lewis's similarity metric requires that we avoid backtracking and that we always foretrack.

miracle occurs which makes the antecedent true. *If, however, the antecedent of the counterfactual is true*, then the formal constraints placed on the similarity relation guarantee that the counterfactual is true just in case the consequent holds in the actual world (Lewis 1973, 560).

Before giving Lewis's simple counterfactual theory, let us stipulate that events can be causes and effects only if they *actually* occur. If c and e are actual events, then the (a)-counterfactual ('if c had occurred then e would have occurred') which partly defines counterfactual dependence will always be true. Therefore, when c and e are actual events, e counterfactually depends on c just in case the (b)-counterfactual ('if c had not occurred then e would not have occurred') holds. Lewis first proposes to analyze causation as follows:

> When c and e are distinct events, 'c causes e' is true just in case (i) c and e actually occur and (ii) ~O(c) □→ ~O(e).

However, this initial formulation is problematic. Consider a case of so-called *early preemption*.[67]

> Assassin poisons Victim's coffee, Victim drinks it and dies. However, if Assassin had not poisoned Victim's coffee, then Backup would have, and Victim would have died anyway.

Intuitively, we judge that Assassin's poisoning of the coffee causes Victim's death, but it is false that Victim would have died if Assassin had not poisoned his coffee. Although Assassin's action cuts off Backup before he can even get started, the mere presence of Backup is enough to break the dependency between Victim's death and Assassin's action.

[67] This example is from Hitchcock (2007).

Although Victim's death fails to be counterfactually dependent upon Assassin's action, it is nonetheless true that Victim's death is counterfactually dependent upon the presence of poison in his coffee. Victim's death is not *directly* counterfactually dependent on Assassin's action but it is *indirectly* through this intermediate event. The death is counterfactually dependent on the presence of poison in Victim's coffee which is counterfactually dependent on Assassin's action. Assassin's action is connected to Victim's death by a "chain" of counterfactual dependencies. Therefore, as a response to cases of early preemption, Lewis's *simple theory* identifies causation with chains of stepwise counterfactual dependence:

> (ST) When c and e are distinct events, 'c causes e' is true just in case (i) c and e actually occur, and (ii) there is a (possibly empty) set of events $\{d_1, d_2, \ldots, d_n\}$ such that $\sim O(c) \mathrel{\Box\!\!\rightarrow} \sim O(d_1)$, $\sim O(d_1) \mathrel{\Box\!\!\rightarrow} \sim O(d_2)$, …, $\sim O(d_n) \mathrel{\Box\!\!\rightarrow} \sim O(e)$.

It is worth mentioning that c is a cause of e *if* e counterfactually depends on c. This kind of direct counterfactual dependence is sufficient but not necessary for causation (Lewis 1973, 563). Hence, causes make a difference to their effects because whether the effect occurs depends (directly or indirectly) on whether the cause occurs.

Section 4.2: A Defense of Counterfactualism

Following Lewis (1973), we can say that an event e *depends nomically* on an event c just in case there is a nonempty set of laws L and a set F of true propositions concerning matters of particular fact such that (L & F) entails the

material conditional O(c) ⊃ O(e).[68] Furthermore, a proposition P_1 is *counterfactually independent* of a proposition P_2 if and only if P_1 holds regardless of whether P_2 holds. According to Lewis, if the propositions in L and F are counterfactually independent of O(c) and (L & F) entails the right material conditional such that the e depends nomically on c, then it follows that e counterfactually depends on c (Lewis 1973, 564). In other words, we can think of the nomic dependencies of e on c as "grounding" or explaining why e counterfactually depends on c. It is in virtue of e's nomic dependence on c that it is true that e would not have occurred had c not occurred. As Lewis remarks, "Often, perhaps always, counterfactual dependences may be thus explained" (Lewis 1973, 564).

Lepore and Loewer (1983) suggest that Lewis's idea of grounding counterfactual dependencies in nomic dependencies is one way to explain why the behaviors of rational agents are counterfactually dependent on their psychologies (Lepore and Loewer 1983, 640 – 641). We think, as seems likely, that in the actual world there is a set of *ceteris paribus* psychological laws which describes the evolution of rational agents. In any particular circumstance, given that the *ceteris paribus* conditions hold for a rational agent S, these psychological laws entail that S's behaviors depend nomically on her psychological profile. If we assume that these laws and their *ceteris paribus* conditions are counterfactually independent of the propositions describing S's psychological profile, then we can conclude that S's

[68] Presumably this holds only if the laws are deterministic. In what follows, I will assume determinism for the sake of the argument, since Lewis's (1986) amendments for probabilistic causation will complicate matters unnecessarily.

behaviors counterfactually depend on her psychological profile. The *ceteris paribus* psychological laws support that S would not have acted as she did had her beliefs, desires, decisions, intentions, etc. not been what they were. By grounding counterfactuals in nomic dependencies, we can justifiably claim that mentalistic counterfactuals like 'Assassin would not have poisoned the coffee had he not wanted to kill Victim' are true.

Recall that the exclusion problem presents an unappealing *dilemma* for the nonreductive physicalist: a choice between mental epiphenomenalism or causal overdetermination. The program I recommend for dealing with the epiphenomenalist horn of the exclusion dilemma starts from the contention that, in the right circumstances, we can truthfully say that a rational agent's behaviors counterfactually depend on her psychology. However, concluding that the mental events constitutive of an agent's psychology are *among the causes* of their behaviors requires a further premise. Although very closely related, the premise I offer is *not* Lewis's simple theory, but instead the following *sufficient condition* for causation[69]:

> (ST*) When c and e are distinct events, 'c is a cause of e' is true if (i) c and e actually occur and (ii) $\sim O(c) \mathrel{\Box\!\!\rightarrow} \sim O(e)$.

As Jonathan Schaffer (2004) has pointed out, (ST*) and the truth of mentalistic counterfactuals suffices to establish mental causation (Schaffer 2004, 240). This, I submit, is a natural and well-motivated way of understanding how the mental *makes a causal difference* in the world.

[69] Like many others, I think cases of late and trumping preemption present insurmountable difficulties for the necessity of counterfactual dependence for causation.

The proposal, however, is not original. Many philosophers concerned with avoiding the epiphenomenalist horn of the exclusion dilemma have suggested that mental causation can be vindicated by appealing to mentalistic counterfactuals. Lynne Rudder Baker (1993) says the question 'Does what we think ever affect what we do?' deserves a simple answer once we reflect on our explanatory practices and the truth of certain counterfactuals (Baker 1993, 93). Tyler Burge (1993) suggests something similar with his remarks that the "difference" the mental makes is "specified by psychological causal explanations, and by counterfactuals associated with these explanations" (Burge 1993, 115). Furthermore, and more along the lines of my proposal above, Terrance Horgan (1997) defends mental causation by appealing to what he considers to be a plausible and well-motivated conception of causation whose "leading idea is that causal properties are ones that figure in robust, objective, patterns of diachronic counterfactual dependence" (Horgan 1997, 179). Barry Loewer (2007) argues that "this much mental causation is near enough to our folk conception of mental causation to underwrite the role of causation in folk psychology, rational deliberation, action theory, and so on" (Loewer 2007, 255). If counterfactual dependence is sufficient for causation, then mental events do make a causal difference to what we do.[70]

[70] Loewer (2001, 2002) suggests something slightly more complicated *vis-à-vis* mental causation: m causes e if ~O(m) □→ ~O(e) and "there is no lower level event that preempts this relation that is not itself preempted by m" (Loewer 2002, 660). A lower level event preempts m with respect to e if e would not have occurred had the lower level event not occurred *and* e would have occurred had the lower level event occurred without m. Loewer (2007) appears to abandon this more complicated proposal. See Section 4.2.1 fn. 71 for further discussion.

Before moving forward, it will be useful to make some distinctions amongst the aforementioned authors and to get more clear on how I understand (ST*). Baker (1993) recommends looking to counterfactual dependence and our explanatory practices to vindicate the causal efficacy of the mental, but she does so in such a way to avoid commitment to *conceptual or metaphysical theses* about causation. Several passages in Burge (1993) also suggest something along these lines. As I interpret them, neither author attempts to ground mental causation by appealing to counterfactual dependence in the exact same way that I do. I believe their project is more guarded: counterfactual dependence can ground our folk-explanatory practices as we consider the behavior of agents. Although counterfactual dependence plays an important role in avoiding the threat of epiphenomenalism for Baker and Burge, I do not wish to lump their positions together with mine. The position I shall defend in this chapter attempts to ground mental causation by appealing to (ST*) which I interpret as being a *metaphysical claim* about causation. I will refer to this less guarded position as *counterfactualism*. Amongst the proponents of counterfactualism, I include authors such as Horgan (1997), Loewer (2001, 2002, 2007), and Schaffer (2004b) where there is *at least* a commitment to (ST*) understood as a metaphysical thesis about causation.[71]

Counterfactualism attempts to vindicate mental causation by appealing to (ST*), a conceptual or metaphysical thesis about causation. What this means is that

[71] Thanks to Neal Judisch for urging that I make the difference between these authors more clear.

counterfactualists take (ST*) to be a very general, and indeed necessary, truth about causation. In other words, (ST*) is not just an empirical generalization about causal relations *in the actual world*, nor is it a generalization about such relations in a wider but still restricted set of worlds (e.g., nomologically possible worlds). Instead, the counterfactualist holds that (ST*) is a truth about causation that holds in all of the metaphysically possible worlds. When the counterfactualist says that (ST*) is a necessary truth, they mean it in an absolute or unrestricted sense of necessity. That is, for any metaphysically possible world *w*, if there is counterfactual dependence between distinct events in *w*, there is causation between those events in *w*. As should be fairly clear, the counterfactualist need not hold that the reverse is the case: they might believe that there are possible worlds in which causation *but no counterfactual dependence* is instantiated between distinct events. I am one of these counterfactualists (see fn. 66). Therefore, I do not hold that counterfactual dependence is somehow conceptually prior or more fundamental than causation, but only that these concepts are tied to one another as specified in (ST*). On the other hand, the counterfactualist does hold that counterfactual dependence is *explanatorily prior* to causation. This is why they claim we are able to ground or explain why mental events are among the causes of bodily effects by appealing to facts about counterfactual dependence.

In the next few sections I shall argue for counterfactualism, the view that (ST*) can adequately ground and thereby vindicate mental causation. However, Jaegwon Kim (1998, 2007) has mounted several arguments against

counterfactualism. He claims that counterfactual dependence is not enough to "vindicate mental causal efficacy and … dissipate our epiphenomenalist worries" (Kim 2007, 236). Kim's first reason centers around his conception of what our philosophical task is in our attempt to vindicate mental causation. If mentalistic counterfactuals are true, they must be made true *by something*. Mentalistic counterfactuals merely mark the surface of some ill-understood phenomenon and the real philosophical task is to understand this phenomenon. The second, and more serious concern, is Kim's claim that the truth of mentalistic counterfactuals is consistent with the mental being utterly non-efficacious (Kim 2007, 234). In other words, Kim argues that the counterfactualist cannot distinguish between genuine causal relations and pseudo-causal relations that hold between mental events and bodily effects. The final problem concerns the connection between agency and mental causation. According to Kim, agency requires mental causation to be a "thick" or productive relation, a kind of relation inconsistent with that recommended by (ST*). This, Kim argues, shows that (ST*) cannot sustain the mental causation required by genuine agency. The first part of my defense of counterfactualism will be to address each of these concerns.

Section 4.2.1: Pseudo-Causal Relations and Mental Causation

Recently, Jaegwon Kim (1998, 2007) has stressed that counterfactual dependence cannot do the work of grounding mental causation. The following passage summarizes nicely Kim's reasons for being skeptical of counterfactualism:

> What the counterfactual theorists need to do is to give an account of just what makes those mind-body counterfactuals we want for

> mental causation true, and show that on that account those counterfactuals we don't want, for example, epiphenomenalist counterfactuals, turn out to be false. Merely to point to the apparent truth, and acceptability, of certain mind-body counterfactuals as a vindication of mind-body causation is to misconstrue the philosophical task at hand (Kim 1998, 71).

The first reason Kim cites is that the counterfactualist leaves unanswered the real question: why are the mentalistic counterfactuals true? As he says, simply asserting that they are "misconstrues" our philosophical task of vindicating mental causation. The second reason is that in some cases the counterfactualist is committed to the presence of mental causation when *ex hypothesi* there is no such causation. For instance, Kim thinks the counterfactualist cannot distinguish between pseudo-causal relations and genuine causal relations. Let us consider each of these objections in turn.

One way of understanding Kim's request for what makes mentalistic counterfactuals true should be satisfied by noting that counterfactuals are *nomically grounded*. Before presenting (ST*), I introduced Lewis's proposal that counterfactuals are explained by nomological dependencies. Roughly, the idea is that, given some further assumptions, we can explain why certain mentalistic counterfactuals are true by appealing to the *ceteris paribus* psychological laws that govern the evolution of rational agents. It is true that Assassin would not have poisoned the coffee had he not wanted to kill Victim *because* in all of the most similar worlds in which Assassin does not want to kill Victim he does not poison the coffee. Whether this is so depends in part on the laws governing the actual world, since world similarity is evaluated *inter alia* in terms of "similarity in

conformity to laws of the actual world" (Loewer 2002, 660). Furthermore, Lewis's proposed similarity metric directs us to evaluate the counterfactuals that define counterfactual dependence by considering those worlds that *match the history of the actual world as much as possible* and that contain *a small, local inconspicuous violation of law* which makes the antecedent true. Hence, what the laws of the actual world are "determines what counts as a 'small violation'" (Loewer 2007, 256). Grounding mentalistic counterfactuals in psychological laws and showing how they follow from such laws explains why these counterfactuals are true.[72]

Nevertheless, I think Kim would be unsatisfied by these remarks. Consider what he says with respect to Fodor's solution to the problem posed by epiphenomenalism:

> To be sure, if there are causal laws in psychology, they will license ascriptions of causal responsibility to psychological properties and ground psychological causal relations. The crucial question unaddressed by Fodor is whether psychological laws *are* causal laws – that is, whether the regularities we observe in the psychological domain are causal regularities (Kim 2007, 232).

And Kim explains that Fodor's neglect of this question results from his "regularist-nomological conception of causation" (Kim 2007, 232). One could easily translate this very complaint against the counterfactualist:

> If there are causal laws in psychology, they will license ascriptions of causal responsibility to psychological properties and ground psychological causal relations. The crucial question unaddressed by the counterfactualist is whether psychological laws *are* causal laws –

[72] Kim would seem to agree with this point when he remarks that "one crucial respect in which the comparative similarity of worlds is to be determined evidently involves the similarity of laws holding in them. It is difficult to see how evaluations of [counterfactual] conditionals … could avoid adverting to laws and regularities" (Kim 2007, 234).

> that is, whether the regularities we observe in the psychological domain are causal regularities.

Presumably, Kim would make a similar complaint that the counterfactualist's neglect of this question results from their counterfactual conception of causation. Now, recall that the counterfactualist wants to explain mental causation in terms of mentalistic counterfactuals and, furthermore, claim that these counterfactuals are grounded in the psychological laws of the actual world. But, for Kim, the real philosophical task is determining whether these psychological laws are *causal laws*. In short, the counterfactualist's appeal to (ST*) to ground mental causation reverses the proper direction of explanation. It is not counterfactuals (and laws) that ground causation, but causation that grounds counterfactuals and laws.

This criticism is difficult to assess because I believe it reflects a deep and fundamental philosophical disagreement about the nature of laws and causation. Yet, on the other hand, Kim's criticism looks to beg the question against the counterfactualist. The counterfactualist project is premised on explaining mental causation in terms of the counterfactuals defining counterfactual dependence. As outlined above, these counterfactuals can be explained by *inter alia* the psychological laws of the actual world. So, Kim's charge that the counterfactualist neglects the *real task* of determining whether these laws are causal looks to be, at bottom, an expression of his dissatisfaction with their project. If certain mentalistic counterfactuals are true and *(ST*) presents a genuine sufficient condition for causation*, Kim's further requests seem inappropriate. But Kim is not without a response here, for he questions whether (ST*) presents a genuine sufficient

condition for causation. Are the psychological laws real, genuine causal regularities or "mere reflections of the causal regularities at a more fundamental level" (Kim 2007, 232)? What he tells us is that there is a difference between "genuine, productive and generative causal processes" and "the noncausal regularities that are observed because they are parasitic on real causal processes" (Kim 1998, 45).

The difficulty, then, for the counterfactualist is *the problem of pseudo-causal relations*. Suppose there are psychological laws and that these laws ground the right kinds of mentalistic counterfactuals. This is still not enough to ground mental causation because the truth of these counterfactuals is *consistent* with there being no real, genuine causation involving mental events at all. Perhaps, as Kim suggests, "the observed regularity arises out of a genuine causal process" at a lower level where all the real causation is happening (Kim 2007, 234). If Kim can show us that this is a real possibility, the kind of "mental causation" the counterfactualist presents to us will be a mere facsimile, a pseudo-causal relation at best, and not the genuine thing. Kim's charge is that (ST*) is committed to mental causation when *ex hypothesi* there is none.

Jill thinks she left her keys on the bookstore counter so she returns to the bookstore. However, Jill's thought is *epiphenomenal* with respect to her returning to the bookstore and it is some neurophysiological event n in Jill's brain that is a cause of both her thought that she left her keys on the bookstore counter and of her

returning to the bookstore. Kim claims that in the circumstances outlined above the following counterfactual is intuitively true (Kim 1998, 71):

> (E) If Jill's thought had not occurred, then she would not have returned to the bookstore.

Presumably, if Jill's thought had not occurred, the neurophysiological event n that caused it would not have occurred and she would not have returned to the bookstore. This is problematic, since *ex hypothesi* Jill's thought is an epiphenomenon, yet (ST*) delivers the verdict that Jill's thought causes her returning to the bookstore.

The counterfactualist has a straightforward reply to Kim's objection: the counterfactual (E) is evaluated incorrectly. Kim's evaluation involves *backtracking*, which Lewis explicitly forbids when he defines counterfactual dependence (see Section 4.1). If we use Lewis's similarity metric, then the set of most similar worlds includes those worlds where Jill's thought does not occur, its cause n still does, and Jill still returns to the bookstore. The most similar worlds involve a "small, local miracle" which violates the causal law connecting n with Jill's thought, but does not interfere with the occurrence of n or the causal law linking it with Jill's returning to the bookstore. Hence, if we evaluate the counterfactual properly, it comes out false. As Loewer (2002) points out, when "evaluating Lewis's counterfactual account of causation, one must employ counterfactuals characterized by Lewis's proposed similarity relation and not merely appeal to intuitions about counterfactuals" (Loewer 2002, 322).

Next, consider a scenario similar to the one above, but this time Jill's thought emerges from, *and is not caused by*, the neurophysiological event n and it is n that causes Jill's returning to the bookstore. Once again, Jill's thought is stipulated to be epiphenomenal with respect to her returning to the bookstore. Again, Kim claims that in these circumstances (E) is intuitively true (Kim 2007, 234). Presumably, if Jill's thought had not occurred, its "neural basal conditions" n would not have occurred and so Jill would not have returned to the bookstore. This poses the same problem for the counterfactualist, since *ex hypothesi* Jill's thought is epiphenomenal with respect to her returning to the bookstore, yet (ST*) licenses the conclusion that the emergent mental event is a cause of her bookstore returning behavior.

The counterfactualist should reply that Kim has once again evaluated the counterfactual (E) incorrectly; although, this time the issue is that Kim gives more weight to avoiding inconspicuous, local violations of law than to maximizing the spatiotemporal region throughout which perfect match of particular fact prevails. The question is which of the following sets of worlds are more similar to the world considered as actual. The first set contains worlds in which Jill's thought does not occur because its neurophysiological base does not occur. In these worlds, Jill does not return to the bookstore. The second set contains worlds in which Jill's thought does not occur because the "emergent law" connecting the neurophysiological base to Jill's thought is violated. In these worlds, the neural basal conditions still occur and so Jill still returns to the bookstore. The first set of worlds matches the world

considered as actual less in *matters of particular fact* than the second set of worlds (i.e., the basal conditions fail to occur in the first set but still occur in the second set). On the other hand, the second set of worlds matches the world considered as actual less in *conformity to its laws* than the first set of worlds (i.e., the emergent law is violated in the second set but not the first).

Now, Kim appears to think that the first set of worlds is more similar to the world considered as actual than the second set of worlds. But if we appeal to Lewis's similarity metric, this is simply mistaken. Lewis (1979) tells us that it is more important to "maximize the spatiotemporal region throughout which perfect match of particular fact prevails" than to avoid small, local violations of law (Lewis 1979, 472). Hence, the second set of worlds with its more extensive match of particular fact and a small violation of the emergent law is more similar to the world considered as actual than the first set of worlds. This means that when evaluated properly according to Lewis's similarity metric, (ST*) derives the correct result that (E) is false. As long as we are careful in our evaluation of the counterfactual (E), neither scenario poses a problem for the counterfactualist.[73]

Although these responses are technically correct, and we must appreciate Loewer's point that evaluating mentalistic counterfactuals must be done by using

[73] Loewer (2001, 2002) offers a slightly different solution to these cases than in his (2007). He takes counterfactual dependence *with some additional requirements* as sufficient for causation (see fn. 70). To avoid the conclusion that (E) is true in either scenario, Loewer holds that the neurophysiological event n *preempts* Jill's thought with respect to her returning to the bookstore. This means that Loewer thinks that (a) Jill would not have returned to the bookstore had the neurophysiological event n not occurred and (b) Jill would have returned to the bookstore had n occurred without Jill's thought. The additional requirements outlined in fn. 70 seem to me unmotivated and a simpler solution is to hold, as Loewer (2007) appears to, that Kim evaluates (E) incorrectly.

Lewis's similarity metric, there is something unpersuasive about these replies.[74] Why, for example, should backtracking be banned in causal contexts? Additionally, what reason is there to place more weight on maximizing match of particular fact than conformity to laws? If all that separates the counterfactualist from the epiphenomenalist is which worlds count as most similar to actuality, it seems as if we are asking Lewis's metric to do *a lot* of heavy duty philosophical work. I cannot deny that there is something to the residual feeling that counterfactualism has merely stipulated mental causation into existence rather than truly vindicating it. Though I cannot develop a full-fledged defense of the Lewis's metric here, I do want to offer a potential diagnosis of why there remains a resistance to it and reiterate some reasons in its favor.

We should recognize that Lewis's similarity metric is not the *only* one available in our everyday evaluation of counterfactuals. It is no part of Lewis's simple theory and (ST*) that the proposed metric which disallows backtracking is the only correct way to evaluate counterfactuals. As is well recognized, counterfactuals are "infected with vagueness" and the different ways of resolving this vagueness involve different metrics to compare similarity amongst worlds. If some of these metrics allow backtracking (which some certainly do) (see, e.g., the example discussed in Section 4.1), this at least partly accounts for why there remains some resistance to the idea that we should disallow backtracking *in causal contexts*. But this, I think, only makes the question of why we should accept

[74] I want to thank Martin Montminy and Neal Judisch for pressing these worries.

Lewis's metric more forceful, since it is not intuitively the most plausible, nor is it the one we always employ in our evaluation of counterfactuals. The answer is that there are good *theoretical reasons* to adopt Lewis's similarity metric especially in causal contexts.

(a) Lewis (1979) notes an asymmetry in counterfactual dependence. The future counterfactually depends on the present and the present on the past, but "it is at best doubtful whether the past depends counterfactually on the present, whether the present depends on the future, and in general whether the way things are earlier depends on the way things will be later" (Lewis 1979, 455). He argues that this asymmetry is a plausible explanation of the asymmetry of causation (i.e., that causes ordinarily precede their effects in time) and the asymmetry in our conception of the future as an open, "multitude of alternative possibilities" and the past as a closed, fixed "actuality". If backtracking is permitted in ordinary contexts (and is not, as Lewis says, only permitted in special contexts), then we no longer have the explanation of these asymmetries at our disposal. The truth of various *backtracking counterfactuals* entails that the past is counterfactually dependent on the present and the present on the future. This is a marked loss in explanatory power.

(b) Bennett (1974) argues that we cannot permit backtracking in ordinary contexts if we are to have "good grounds" for believing in the truth of counterfactuals. Suppose the world is deterministic. If we allow backtracking, then any counterfactual antecedent entails "an earlier difference which will imply a

still earlier one which … and so on back for a million years" (Bennett 1974, 391). Furthermore, if we trace these earlier differences in the past forward through time in accordance with the deterministic laws, we end up with a very different future and likely one we have no good idea about. As Bennett says, the point is "we cannot do this" and "because we cannot do it, we adopt [similarity] standards which don't require us to do it" (Bennett 1974, 391). Our similarity metric permits us to make counterfactual suppositions about the present while *holding the past fixed as much as possible*, that is, we disallow backtracking. Note that the assumption of determinism is crucial in Bennett's reasoning above. If indeterminism is true, then a counterfactual antecedent does not entail a difference in the past and so on back a million years. But surely it would be absurd to suppose that we can only have good grounds for believing in the truth of counterfactuals if *indeterminism* is true.

(c) Lewis's simple theory and (ST*) have seemed to many to be a plausible and well-motivated analysis of causation. But if we permitted backtracking in causal contexts neither analysis gets even the simplest cases of causation correct.[75] If our counterfactual supposition is that some effect e had not occurred, this entails that its cause c would have to have not occurred. The truth of *backtracking counterfactuals* such as 'if e had not occurred then c would not have occurred' means that an effect e is a cause of its cause c. This absurd result places in jeopardy the simple theory's and (ST*)'s claim to be *at the very least* a plausible

[75] See (Hall 2004, 233 – 234) for a similar point.

analysis of causation. So unless the simple theory and (ST*) are just obviously mistaken, there *must* be a ban on backtracking in causal contexts.

(d) Lewis's non-backtracking similarity metric reflects an important connection between causation and manipulation.[76] The idea is that changing or intervening on a cause is a means to or effective strategy for manipulating its effects. If an intervention took place changing whether some event occurs, this should influence the manner in which its effects occur or perhaps even whether its effects occur at all. When Lewis's non-backtracking metric instructs us to hold the past fixed as much as possible and posit *a small, inconspicuous miracle* making our counterfactual supposition true, this miracle corresponds roughly to a potential intervention made on a cause to manipulate its effects.

(e) The idea of holding the past fixed as much as is possible and positing a "simple, localized, inconspicuous" miracle has an analogue in experimental design and is reflected in the reasoning that underlies experiments meant to discover causal relationships.[77]

Suppose a scientist wants to discover the relationship, if any, that holds between some variable V_1 and another variable V_2. In the simplest kind of experiment, the scientist controls the value of the independent variable V_1 and observes changes, if any, that manifest in the value of the dependent variable V_2. If, under the conditions of the experiment, changes in the value of V_1 are correlated with changes in the value of V_2, then usually an inference is made that values of V_1

[76] Some authors have argued this connection is conceptual (see, e.g., Woodward 2003).
[77] See (Menzies 2003, 151) for a similar point.

make a causal difference to the values V_2. However, this inference from correlation to causation can be undermined if it is shown that the experimental manipulations performed by the scientist on the independent variable involved some confounding factor (i.e., a "gratuitous" departure from the control conditions) that could influence the value of the dependent variable. The ideal experiment is one in which changes in the independent variable involves *only the changes necessary* to determine if there is some change in the dependent variable. If experiments designed to discover causal relationships are properly conducted, they should hold fixed as much as they possibly can, introduce a "small, localized" change somewhere in the experimental conditions, and determine if this is correlated with a change in the dependent variable.[78]

There is a striking similarity here to what Lewis's non-backtracking metric instructs us to do when evaluating the counterfactuals relevant in causal contexts. Imagine the actual world as the *control conditions* in which distinct events c and e occur. According to Lewis's metric, the set of most similar worlds are the *experimental conditions* in which we hold the past fixed as much as is possible, posit a small, local miracle such that the event c does not occur. We then make a "difference observation" to determine if the event e occurs. If it does not occur, then the simple theory and (ST*) license the conclusion that c *makes a causal difference* to e. If we allow backtracking, then our counterfactual supposition that c

[78] As Menzies (2003) points out, this is the idea behind J.S. Mill's method of difference which involves "a difference observation between a positive instance in which some effect E is present and a negative instance in which E is absent. If some condition C is present in the positive instance and absent in the negative instance, it is, at least, part of what makes the difference to E" (Menzies 2003, 151).

does not occur would entail that some earlier event(s) would had to have not occurred which has the potential to influence whether the event e occurs. Lewis's ban on backtracking in causal contexts is analogous to the scientist's attempt to minimize the presence of confounding factors in experimental design.

For these reasons, I believe there is strong epistemic pressure to accept Lewis's non-backtracking similarity metric, which gives more weight to maximizing the region of match of particular fact then avoiding violations of law.

Yet there remains the residual feeling that counterfactualism presents us with an inadequate account of mental causation. I believe this feeling stems partly from the intuition that the kind of *causation*, the kind of *mental causation*, we are ultimately left with is pretty "thin". If it is causation at all, it is "causation lite" and lacks the substance of the real thing. This seems to amount to the position that counterfactual dependence really isn't sufficient for causation after all, since (ST*) cannot guarantee that causes *produce* or *generate* their effects. As Elizabeth Anscombe writes:

> Causality consists in the derivativeness of an effect from its cause. This is the core, the common feature, of causality in its various kinds. Effects derive from, arise out of, come of, their causes. For example, everyone will grant that physical parenthood is a causal relation (Anscombe 1993, 91 – 92).

Counterfactualism simply does not present us with a robust enough conception of causation to satisfy our intuitions that causes produce and generate their effects, that causes are "something from which the effects derive their existence and occurrence" (Kim 2007, 235). In the next section, we will see that the intuition that

causation is a productive or generative relation has a prominent role in Kim's final argument against counterfactualism. I shall argue that his argument falls short of its intended conclusion and then, in Section 4.2.3, I address this *production intuition* more directly.

Section 4.2.2: Agency and Mental Causation

The most recent problem raised by Kim (2007) against the counterfactualist concerns the connection between agency and mental causation. He argues that agency requires "thick" or productive causation in which "causes are connected to their effects via spatiotemporally continuous sequences of causal intermediaries" (Hall 2004b, 225). Therefore, our account of mental causation must satisfy the strictures placed on the causal relations involved in an agent's *bringing things about* in the world. He remarks,

> It seems to me that mere counterfactual dependence is not enough to sustain the causal relation involved in our idea of acting upon the natural course of events and bringing about changes so as to actualize what we desire and intend (Kim 2007, 236).

I could not agree more with Kim that "we care about mental causation because we care about human agency" (Kim 2007, 236). Cursory reflection reveals that agency requires among other things the causal efficacy of a wide variety of mental phenomena (beliefs, desires, intentions, choices, decisions, etc.). If such mental phenomena ended up, as it were, *making no difference* to the ways in which our bodies move, then the conception of ourselves as genuine agents – actors in the world who bring things about – would be swept right from underneath us.

However, notwithstanding our agreement on that point, Kim argues the connection between mental causation and agency has troublesome consequences for counterfactualism. He reasons as follows:

> An agent is someone who, on account of her beliefs, desires, emotions, intentions, and the like, has the capacity to perform actions in the physical world: that is, to cause her limbs and other bodily parts (e.g., vocal cords) to move in appropriate ways so as to bring about changes in the arrangement of objects and events around her – open a door, pick up the morning paper, and make a cup of coffee. It seems to me that without productive causation, *which respects the locality/contiguity condition*, such causal processes are not possible (my emphasis) (Kim 2007, 236).

What Kim is saying here is that real, genuine agency is *not possible* without causation that guarantees mental events to be connected to their effects by *spatiotemporally local and contiguous processes*. He elaborates that these processes are "constituted by phenomena such as energy flow and momentum transfer, an actual movement of some (conserved) quantity" (Kim 2007, 236). Of course, mere counterfactual dependence between events does not guarantee that the events are connected in this way and, in fact, the events need not be *connected* at all.[79] The conclusion, then, is that human agency requires something more than

---

[79] Cases of "double prevention" make this point clear:

> Suzy is piloting a bomber on a mission to blow up an enemy target, and Billy is piloting a fighter as her lone escort. Along comes an enemy fighter plane, piloted by Enemy. Sharp-eyed Billy spots Enemy, zooms in, pulls the trigger, and Enemy's plane goes down in flames. Suzy's mission is undisturbed, and the bombing takes place as planned. If Billy hadn't pulled the trigger, Enemy would have eluded him and shot down Suzy, and the bombing would not have happened (Hall 2004, 241).

We can add that Billy's shooting down of Enemy took place in a region of space far removed from Suzy's bombing of the target. Presumably, this helps to make it more clear that no local and contiguous process connects the events. So, although the bombing is counterfactually dependent on

counterfactual dependence between mental events and their effects; only "thick", productive mental causation will do.

The question that Kim leaves unanswered is why human agency is not possible without productive mental causation. The answer appears to be that such a notion of mental causation is a part of *our very concept of an agent*, that is, we know *a priori* that the mental causation involved in agency involves spatiotemporally local and contiguous processes connecting cause and effect. That this is the sense in which human agency is not possible without "thick" mental causation is suggested by Kim's remarks that "mere counterfactual dependence is not enough to sustain the causal relation *involved in our idea of acting upon the natural course of events*" (my emphasis) (Kim 2007, 236). Assuming that Kim's view is that agency requires productive mental causation as a matter of conceptual necessity, I find his claim highly dubious. Many religious traditions countenance the genuine possibility of wholly disembodied agencies and, given that the concept of such an agency plays an important role in many human lives, there is *prima facie* reason to think that it is coherent. For instance, the God of the Abrahamic traditions is thought of as an agency whose volitions not only brought into existence the physical universe as a whole, but also make a difference to the physical events that happen therein. Yet the idea of the Abrahamic God involves an entirely non-physical entity whose volitions could not be connected to their

---

Billy's action, there is no *spatiotemporal connection* between them. See (Hall 2004a, 2004b) for a detailed discussion of these kinds of cases.

effects via any spatiotemporally local and contiguous processes.[80] A wholly disembodied agency like the Abrahamic God is *prima facie* a genuine conceptual possibility. If this is right, then Kim's claim that from *the very concept of an agent* we know that mental events are connected to their effects via spatiotemporally local and contiguous processes is mistaken.

Furthermore, there is a way of understanding a suggestion by Loewer (2007) which also undermines Kim's claim about agency requiring productive mental causation. Loewer asks us to suppose that

> The batteries of counterfactuals that are associated with volitional control of bodily movement, with stimuli and perceptual belief, with rational thinking, and so on obtain but without the transfer of energy and without productive causation connecting individual events. Perhaps this would be the situation, if, as Jonathan Edwards[81] seemed to think, one state of the universe doesn't produce the next via law but rather the states are produced one after another by God in a manner of a movie projector … Would we stop taking aspirin for headaches, cease taking seriously the readings on thermometers, and so on? Would we think that causation as dependence (without production) is not worth having? (Loewer 2007, 258 – 259).

Intuitively, we think of Loewer's world without "thick" mental causation as a world where there is still genuine agency, since humans are still able to exercise volitional control over their behaviors. My decision to take an aspirin for headaches is still something *that I do*, it is still something *under my control*, even

[80] I do not mean to suggest that God's volitional action on the physical world is entirely unproblematic. I only point out that the concept of a disembodied agency like God does not involve a contradiction and so the idea of having an agency which influences the world without a spatiotemporally contiguous process connecting cause and effect is not contradictory.

[81] In a discussion of persistence through time, Edwards tells us that "the existence of created substance, in each successive moment" is "wholly the effect of God's immediate power, in that moment, without any dependence on prior existence, as much as the first creation out of nothing". See *Jonathan Edwards*, ed. C.H. Faust and T.H. Johnson (New York, 1935): 335.

though no local and contiguous process connects my decision to my taking of the aspirin. The claim that there is no real agency in this world, no real mental causation, is hard to swallow given that it remains true that what I decide *makes a difference* to what I do. I wouldn't have taken the aspirin and been relieved of my headache if I hadn't decided to take it. If this is the right way to describe such a possibility, then again, *pace* Kim, agency does not require, in the relevant sense, a local and contiguous process connecting mental events with their effects.

As I have tried to make clear, Kim openly proceeds from *conceptual considerations*; it is from our very *concept* of an agent that it is claimed we know mental causes must be connected to their effects via local and contiguous processes. Hence, Kim's own argumentative strategy against the counterfactualist is vulnerable to the conceptual possibilities outlined above. His thesis is false if it is understood as giving conceptual conditions for agency. But perhaps it remains true if it is understood instead as a claim about the way agency is realized in *the actual world and nomically similar ones*. Specifically, this modified form of Kim's thesis states that the mental causation sustaining agency *in this restricted set of worlds* involves mental events which are connected to their effects via spatiotemporally local and contiguous processes.[82] This thesis is consistent with the conceptual possibility of disembodied agencies, like the Abrahamic God, and Loewer's world and so is impervious to the aforementioned objections.

[82] This modified thesis will likely seem plausible to those philosophers who eschew a "conceptual analysis" of causation for an *empirical characterization* of what actual world relations are causal. See, for example, (Dowe 2000, 2004).

More importantly, an argument similar to the one initially raised by Kim can still be employed against the counterfactualist. Consider the following: *in the actual world*, agency is not possible without causation that guarantees mental events to be connected to their effects by spatiotemporally local and contiguous processes; (ST*) cannot guarantee that *in the actual world* mental events are connected to their bodily effects in this way; therefore, agency, at least as it is realized in the actual world, requires something more than mere counterfactual dependence. Without a doubt, counterfactualism should be rejected if the causation recommended by (ST*) cannot sustain agency as it is actually realized. Luckily for the counterfactualist, the kind of productive mental causation referenced by Kim's thesis is inconsistent with the way in which mental causal relations are implemented in the human organism. In the next section, I argue for this claim and discuss its consequences for avoiding the epiphenomenalist horn of the exclusion dilemma.[83]

If my objections in this section are on the mark, I believe there is sufficient reason to reject the initial formulation of Kim's final argument against counterfactualism. However, we might still wonder exactly what real, genuine agency "requires" in an absolute and unrestricted sense of that term. The answer, I believe, is Loewer's idea of *volitional control*: agency is impossible without the capacity to exercise the right kind of volitional control over one's bodily movements. As an illustration of this idea, consider the important and influential

[83] See (Schaffer 2000a, 288 – 289) for a brief discussion of this kind.

theory developed by Fischer and Ravizza (1998) in which the agency required for moral responsibility is understood in terms of *guidance control*. An agent exhibits guidance control of some action when the action issues from the agent's own moderately reasons-responsive mechanism. Roughly, Fischer and Ravizza argue that a moderately reasons-responsive mechanism is to be understood as a mechanism that displays an appropriate sort of *receptivity* and *reactivity* to reasons. A mechanism with the appropriate sort of receptivity is such that there is some (nomologically) possible world in which the same kind of mechanism operates in which the agent would recognize something as a sufficient reason to do other than she actually did. Furthermore, the mechanism is disposed to respond to a *regular and understandable pattern* of actual and hypothetical reasons, some of which are moral reasons (Fischer and Ravizza 1998, 68 – 73). A mechanism with the appropriate sort of reactivity is such that there is some (nomologically) possible world such that the same kind of mechanism operates, the agent recognizes there is a sufficient reason to do otherwise, and the agent does otherwise *for that reason* (Fischer and Ravizza 1998, 73 – 76).

Importantly, the agent's mechanism must also be the agent's "own" in the sense that the agent has in the past, and most likely as a result of their moral education, *taken responsibility for* acting from that particular kind of mechanism (e.g., the mechanism of practical reason) (Fischer and Ravizza 1998, 215).[84] They explain,

[84] Fischer and Ravizza point out that practical reason is not the only kind of mechanism an agent can take responsibility for and so is not the only mechanism which grounds the agent's moral

> First, an agent must view himself – when acting from certain mechanisms – as an agent; he must see that certain upshots in the world are the results of his choices and actions. Second, an agent must view himself as an apt target for the reactive attitudes … Finally, the cluster of beliefs specified by the first two conditions must be based, in an appropriate way, on the individual's evidence (Fischer and Ravizza 1998, 238).

In short, an agent's taking responsibility for a mechanism of a certain kind consists in having a cluster of evidentially grounded beliefs about oneself and actions that stem from mechanisms of that kind.

Although this is only a very rough characterization, it is straightforward that Fischer and Ravizza's theory involves the causal efficacy of mental events. However, central to our present purposes, their analysis of both moderate-reasons responsiveness and taking responsibility does not explicitly demand that events be connected to their effects via *spatiotemporally local and contiguous processes*. It appears entirely irrelevant to the agency that sustains responsibility that the events constitutive of the agential mechanisms be connected by spatiotemporally local and contiguous processes or transfer to one another some conserved physical quantity. While Fischer and Ravizza's theory remains controversial, it is a plausible and influential take on the analysis of the agency required for moral responsibility, a take that is *prima facie* entirely consistent with the counterfactualist account of mental causation.

My discussion in the last two sections leads me to conclude that Kim falls short of providing reasonable grounds for his claim that (ST*) cannot ground

---

responsibility. Other non-reflective kinds of mechanisms (e.g., habit) can issue in actions for which the agent can be properly held morally responsible (see (Fischer and Ravizza 1998, 46 – 51, 214 – 215)).

mental causation. There are well-motivated reasons from which the counterfactualist can distinguish genuine mental causation from pseudo-causal relations. Additionally, there is *prima facie* reason to doubt Kim's initial thesis that agency requires mental causes to be connected to their effects via spatiotemporally local and contiguous processes. At best, agency demands the exercise of a certain kind of volitional control. Moreover, if we find it plausible to understand this idea along the lines of Fischer and Ravizza's (1998) analysis of guidance control, agency can be sustained by the kind of causation recommended by (ST*).

Section 4.2.3: The Price of Mental Causation

In Sections 4.2.1 and 4.2.2, I have been primarily concerned with answering the criticisms of counterfactualism raised by Kim (1998, 2007). Despite the shortcomings of these criticisms, I believe there remains a stubborn resistance to counterfactualism as an adequate solution to the epiphenomenalist horn of the exclusion dilemma. The issue, at its most basic level, is that something just seems to be missing from the picture of mental causal efficacy recommended by the counterfactualist. Above I called this the *production intuition*: causation is a productive or generative relation between events. The problem for counterfactualism is that a true vindication of mental causation should reveal it to be a productive or generative relation. The production intuition motivates much of Kim's skepticism toward counterfactualism, since counterfactual dependence does not guarantee mental causes are productive causes. However, in this section, I would like to show that the counterfactualist can go on the offensive against this

production intuition. I shall argue that the price of vindicating mental causation, *at least in the human organism*, is abandoning causation as a productive relation. My argument relies on empirical claims about the physiological mechanisms of human action used by Schaffer (2000a, 2004a) in his defense of causation by disconnection. I believe the conclusions drawn here tell us something deeply important about our approach to vindicating the causal efficacy of our minds.

When c is causally related to e, there is some "mechanism" or underlying structure which we can say *implements* this causal relation. These causal mechanisms have traditionally been illustrated using neuron diagrams with the following conventions: dark circles represent firing neurons and occurring events; empty circles represent non-firing neurons and absences; lines headed with arrows represent a stimulatory connection between neurons and a causal relation between events. Here is a simple example illustrating these conventions in which Terrorist's pressing of the detonator causes an explosion of a bomb.

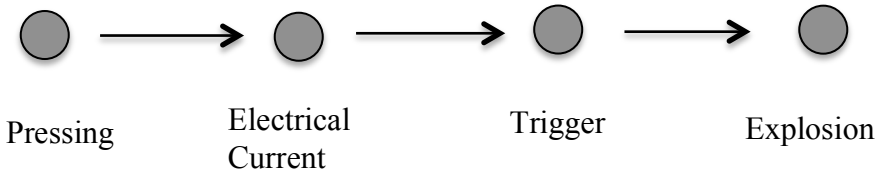

Figure 1: Simple Causal Mechanism

Figure 1 reveals the underlying structure which implements this causal relation: the pressing generates an electrical current which triggers the fuse and leads to the bomb's explosion. What this simple causal mechanism reveals is that the pressing is *physically connected* to the explosion. There are no absences, breaks, or

disconnections in the causal chain running from the pressing of the detonator to the explosion, that is, a spatiotemporally local and contiguous process connects the cause with its effect.

But intuitively not all causal relations are implemented in this fashion. We need to introduce another convention to illustrate this. Lines headed with dots represent *inhibitory connections* between neurons. These inhibitory connections depict one event preventing another from occurring. Inhibitory connections are standardly understood to take priority over stimulatory connections in the sense that if two neurons are connected by a stimulatory connection, the second fires if the first does so long as no other neuron inhibits it.[85] Now, consider a case of "double prevention" given by Ned Hall (2004b), mentioned previously (in fn. 76), in which Suzy is piloting a bomber on a mission to blow up an enemy target.

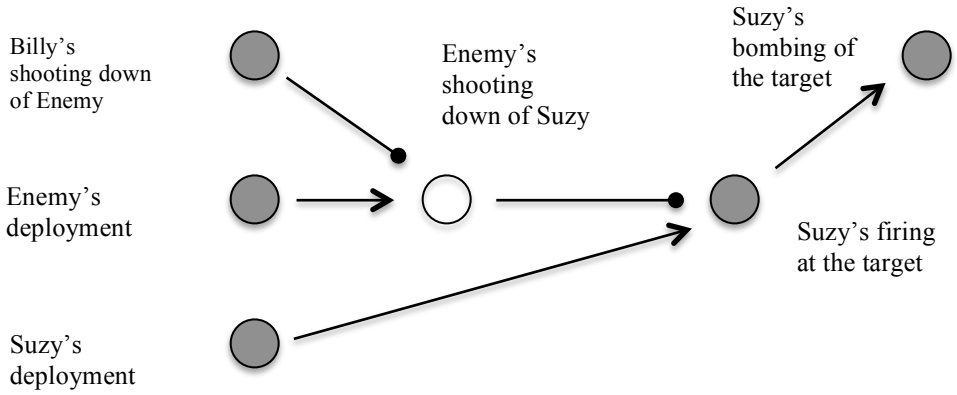

Figure 2: Standard Case of "Double Prevention"

Figure 2 reveals that Billy's shooting down of Enemy *prevented* Enemy from shooting down Suzy, which *would have prevented* Suzy from firing at and thus

[85] I follow (Schaffer 2004a, 197, fn. 1) here.

blowing up the target. Suzy would not have succeeded in her mission of blowing up the target had Billy not shot down Enemy. Hence, according to (ST*), Billy's action is amongst the causes of Suzy's bombing of the target. However, unlike the events depicted in Figure 1, there is no spatiotemporally local and contiguous process connecting Billy's shooting down of Enemy and Suzy's bombing of the target. First of all, an *absence* is intermediate between Billy's action and Suzy's bombing of the target. This absence precludes the former event from being *contiguously connected* with the latter event.[86] Additionally, we can suppose that the interaction between Billy and Enemy takes place in a region of space far away from Suzy implying that the cause is not *locally connected* to its effect.

Whether (ST*)'s verdict is correct here is a complex and controversial matter in the causation literature. Schaffer (2000a, 2004a) has argued that (ST*) gets the right answer, since most of the conceptual connotations of causation are in full force.[87] For example, Billy's action is *statistically relevant* to the bombing, it is *predictive evidence* for the bombing, that the target is bombed is *retrodictive evidence* of Billy's action, and Billy's action is an *effective way* to manipulate whether the bombing occurs. Furthermore, as Hall himself remarks, "Wouldn't we give Billy part of the *credit* for the success of the mission? Isn't Billy's action part of the *explanation* for that success? ..." (Hall 2004b, 242). Nevertheless, Hall (2004b) argues that (ST*) gets the wrong answer:

---

[86] Hall (2004b) considers whether Billy's action *is* connected to Suzy's bombing of the target via a contiguous chain of events that consist in part of *omissions*. However, the prospects of making this work seem slim (see (Hall 2004b, 243) for criticisms of this response).

[87] See (Schaffer 2000a, 2000b, 2004a) for this kind of argument.

> Here [Suzy] is, in one region, flying her plane on the way to her bombing mission. Here Billy and Enemy are, in an entirely separate region, acting out their fateful drama. Intuitively, it seems entirely unexceptionable to claim that the events in the second region have no causal connection to the events in the first – for isn't it plain that no *physical connection* unites them? (original emphasis) (Hall 2004b, 242).

Central to his case is an appeal to the *production intuition*: causes are connected to their effects by way of spatiotemporally local and contiguous processes and Billy's action is *not* connected in this way to the bombing.

Cases of "double prevention", like the one above, are a focal point of disagreement between those who maintain the production intuition and those who accept (ST*), since they represent circumstances where there is counterfactual dependence *but no physical connection* between distinct events. In the rest of this section, I want to show that the causal relations between mental events and bodily effects in the human organism are implemented in a way that is *structurally isomorphic* to cases of "double prevention". Hence, if the production intuition leads one to reject double preventers like Billy's action as genuine causes, then it also leads one to reject mental events as genuine causes of action. This demonstrates that the production intuition comes into conflict with some of the most paradigmatic cases of causation. The price of mental causation is deserting the production intuition and forsaking causation as a productive or generative relation.

First, let us follow Schaffer (2000a, 2004a) and point out that many intuitive cases of causation are structurally isomorphic to the case outlined above.

For example, we have a strong intuition that Assassin's firing of the bullet through Victim's heart is among the causes of Victim's death, but

> Heart piercings cause death only by *disconnection*. The brain is kept alive by an influx of oxygenated blood, and heart piercings cause death by disconnecting this influx, allowing oxygen starvation to run its course (my emphasis) (Schaffer 2000a, 286).

Here is the corresponding mechanism that implements this causal relation:

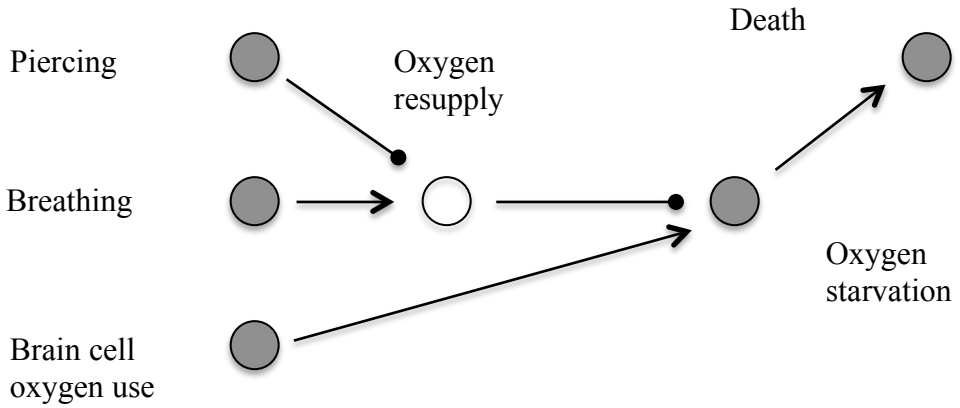

Figure 3: Another Case of "Double Prevention"

A cursory comparison of Figure 2 and 3 reveals that the very same "double prevention" structure implements the intuitively causal relation between the piercing of the heart and Victim's death: the piercing *prevents* the resupply of oxygen which *would have prevented* oxygen starvation. For the same reasons that Billy's action is not connected to Suzy's bombing of the target, the piercing of the heart is not connected to Victim's death. Instead, Assassin's firing of the bullet interferes with the process of oxygen resupply that was keeping Victim alive. The causation here works by *disconnecting a process*, not by a *connecting process*. The price of maintaining that Assassin's action is among the causes of Victim's death is abandoning the production intuition.

This same issue repeats itself when we consider the physiological mechanisms involved in bodily movement. Consider the following description of the "sliding filament theory" of muscle contraction.

> The functional unit of muscle contraction is called the *sarcomere* which is composed of a thick filament called *myosin* and a thin filament called *actin*. The heads of the myosin filaments "want" to bind to sites on the actin, but are blocked by the presence of *tropomyosin* which lays across the actin filament preventing the myosin heads from binding to the actin. When nerve signals in the motor cortex fire, calcium ions (Ca^{2+}) stored in the muscle fiber are released which bind to a structure on the tropomyosin called *troponin*. When the troponin is exposed to calcium ions, it causes the tropomyosin to undergo a "conformational change" moving it away from the actin. This enables the myosin heads to bind to the sites on the actin filament. When the myosin heads bind to the sites on the actin, they "pull" the actin filament resulting in a sliding motion. The sliding of the myosin and actin filaments across one another constitutes the contraction of the muscles.[88]

What we see is that the physiological mechanism of muscle contraction has a "double prevention" structure. Schaffer (2004a) explains,

> Nerve signals only cause muscle contractions … by [disconnection]: the firing of the nerve causes a calcium cascade through the muscle fiber, which causes calcium-troponin binding, which causes *the removal of* tropomyosin from the binding sites on the actin, which causes myosin-actin binding, and thereby causes the actin to be pulled in and the muscle to contract (my emphasis) (Schaffer 2004a, 200).

The neuron diagram of this physiological mechanism makes this "double prevention" structure explicit:

[88] The following YouTube video nicely illustrates the details of the "sliding filament theory": http://www.youtube.com/watch?feature=fvwp&NR=1&v=f0mDFP7qn1Y

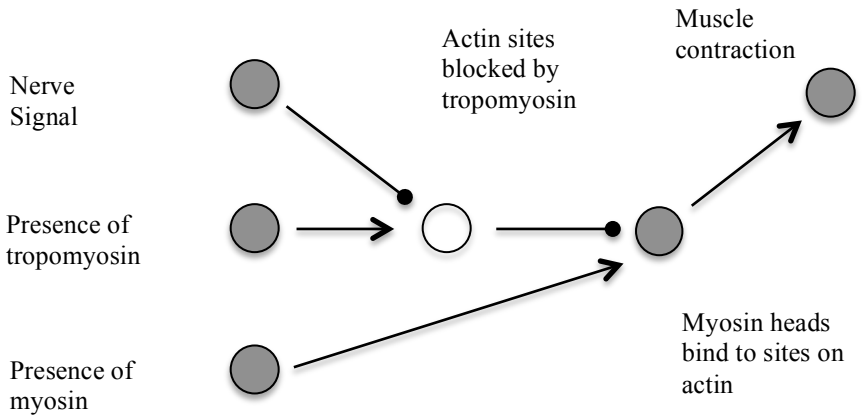

Figure 4: Mechanism of Muscle Contraction

The importance of this empirical claim cannot be understated: *if the "sliding filament theory" of muscle contraction is correct*, then neurophysiological events are not physically connected to their bodily effects by way of spatiotemporally local and contiguous processes. The movement of our bodies results from the complex nerve signals in our brains triggering a biochemical process that disconnects an ongoing physiological process in our muscles. Again, we see that the causation works here by *disconnecting a process*, not by a *connecting process*.

Now, let us incorporate this empirical information into our treatment of the epiphenomenalist horn of the exclusion dilemma. Anyone concerned with vindicating the mental causation of action must admit that the bodily effects standardly attributed to mental events are brought about by way of the movement of the human body, which involves the contraction of muscles. In other words, the causal relation between mental events and their bodily effects is implemented

partly by the sliding filament mechanism depicted in Figure 4.[89] How else could we execute our agential capacities except by way of the physiological mechanisms of the human organism? What this means is that at least part of the mechanism implementing mental causal relations in the human organism exemplifies a "double prevention" structure from which it follows that mental events are *not* connected to their bodily effects via spatiotemporally local and contiguous processes. Mental causation works in part by *disconnecting an ongoing physiological process* in the human body. What this empirical information implies is that the price of holding onto mental causation is giving up the intuition that it is a productive or generative relation. As Schaffer himself notes, "Since all *voluntary human actions* are due to muscle contractions, it follows that all voluntary human actions (perhaps the most paradigmatic of all causes) involve [causation by disconnection]" (Schaffer 2004a, 200).[90]

---

[89] Kim (1993b) seems to admit this much: "From what we know about the physiology of limb movement, we must believe that if the sensation causes my hand to withdraw, the causal chain from the pain to the limb motion must somehow make use of the causal chain from an appropriate central neural event to the muscle contraction; it makes no sense to think that there was an independent, perhaps telekinetic, causal path from the pain to the limb movement" (Kim 1993b, 281).

[90] The above considerations should not be taken as an argument for the claim that the implementing mechanism of mental causal relations in the human organism *must* involve a "double prevention" structure. It is logically possible that mental events cause voluntary human actions, that all voluntary human actions are due to muscle contractions, that all muscle contractions involve causation by disconnection, and yet that mental causation of voluntary human actions does *not* involve causation by disconnection. In other words, it is consistent with the empirical facts outlined above that mental causal relations are implemented in a simple, direct, and productive way; a way independent from how neurophysiological causal relations are implemented. First, Kim rejects that mental causal relations are implemented independently of neurophysiological causal relations (see fn. 89), so the above considerations can be understood as an *ad hominem* against Kim. Second, I do not find it plausible that any *physicalist* – reductive or nonreductive – would accept that mental causal relations are implemented independently of neurophysiological causal relations. Thus, in the context of assuming a physicalist metaphysics of mind, it is plausible that mental causal relations are implemented in a way that involves causation by disconnection. I want to thank Reinaldo Elugardo for bringing this concern to my attention.

Any attempt to vindicate mental causation, at least mental causation in the human organism, demands that one abandon the production intuition. If causation is a productive or generative relation, then causes are connected to their effects by way of spatiotemporally local and contiguous processes. But, as we saw above, mental causes are not connected to their bodily effects in this way. Therefore, causation is *not* a productive or generative relation. Alternatively, we can say that the way in which the causal relations between mental events and their bodily effects are implemented in the human organism is *inconsistent* with productive causation. To my mind, this reveals that the production intuition is a bad, misleading intuition about causation and the stubborn resistance to counterfactualism based on it is equally misleading. *If one accepts mental causation, then one has to reject that causation is a productive or generative relation*. Mental events are not physically connected to their bodily effects. The price of mental causation is abandoning the production intuition.

Near the end of his critique of counterfactualism, Kim urges that a serious commitment to productive causation leads to *reductionism* about the mental. He asks,

> But if we understand causation in mental causation in the productive/generative sense, wouldn't that rule out mental causation – in particular mental-physical causation – too quickly, without any need for an argument? Especially if we require that causation requires energy flow or momentum transfer, how could there be such a process from a mental entity to a physical entity, or in the converse direction? ... Don't all such conceptions of causation, conceptions that require some "real" connections between cause and effect, automatically rule out mental-physical causation (and hence human agency)? (Kim 2007, 239).

These questions betray Kim's position that productive causation does not appear to sit very comfortably with nonreductionism. As he suggests throughout the discussion, this set of views either has "an epiphenomenalist implication" or leads to "the problem of overdetermination" (Kim 2007, 239).

But what Kim fails to realize is that reductionism does not fit comfortably with productive causation either. A consequence of the "sliding filament theory" of muscle contraction is that the neurophysiological events of the human animal are *not* physically connected to their bodily effects. Specifically, given that the contraction of the muscles involves disconnecting the process that keeps them relaxed, neurophysiological events are not connected to their bodily effects by spatiotemporally local and contiguous processes. So, Kim's move to reductionism on the basis of a commitment to productive causation is misguided. *If one is committed to neurophysiological causation, then one has to reject that causation is a productive or generative relation*. The manner in which the human brain is "wired" to the physiology of the human body means that our neurophysiology does not *produce* the movement of our bodies. Surprisingly, the price of neurophysiological causation is giving up the production intuition.

Finally, recall that in Section 4.2.2 we saw two very similar objections to counterfactualism based on two interpretations of Kim's thesis that agency requires productive mental causation. If the "sliding filament theory" of muscle contraction is correct, there is no productive mental causation in the human organism. Therefore, Kim's thesis entails that *there are no human agents*! This is more than

just a highly problematic conclusion; it is a *reductio* of Kim's insistence that genuine agency requires productive mental causation. For these reasons we should reject Kim's final argument against counterfactualism. Furthermore, we can draw similar conclusions here as we did above: *if one accepts that there are human agents, then one has to reject that causation is a productive or generative relation*. At least in the human organism, our agential capacities are exercised in a non-productive, non-generative fashion. Again, the price of human agency is giving up the production intuition.

Before moving forward, I would like to consider a striking admission Kim makes in his discussion of agency, which offers a potential avenue of response on behalf of the production intuition. He remarks that it is possible that at the basic physical level a counterfactual account of causation is correct, since at the "bottom level" this is all we can get (Kim 2007, 232). This claim is *prima facie* in tension with his thesis that agency is (conceptually) impossible without productive mental causation. How is it that agency is *impossible* without productive mental causation and yet there *might be* causation in the actual world that is not productive? The only way to square these two claims is to interpret Kim as implicitly assuming that whatever causation is like at the basic physical level, *mental* causation is of an entirely different sort.

This closely resembles the position defended by Hall (2004a, 2004b) that there are two distinct types of causation: there is productive causation where causes bear a "real" connection to their effects and there is counterfactual dependence

which demands no such thing[91]. If we suppose Hall's "dual causation" thesis is correct, then the production intuition is at least not entirely misleading, since it is true of *one kind of causation*. But it should be fairly clear that this distinction amongst types of causation cannot help Kim maintain his position that *mental causation* is a productive relation. Even if there really is a productive and non-productive kind of causation, we know that the empirical details of human physiology ensure that mental causation does not fall under the productive type. As far as I can tell, the only way to preserve the production intuition in its entirety without sacrificing mental causation is to reject the "sliding filament theory" of muscle contraction.

Section 4.3: The Case for Counterfactualism, and Some Unresolved Issues

The upshot of Section 4.2.3 is that anyone concerned with vindicating mental causation, neurophysiological causation, or human agency had better relinquish the production intuition. Instead, they are better served to search for an account of causation which makes *no demand* that causes be connected to their effects by spatiotemporally local and contiguous processes. Of course, none of this entails that counterfactualism is true or remains our only plausible option for grounding mental causation. Perhaps there are a variety of accounts of causation

[91] Hall's thesis suggests that one might escape the exclusion dilemma by denying the Homogeneity Assumption (see Section 2.2). If mental causation is simply a different kind of causation than neurophysiological causation, our worries about epiphenomenalism and overdetermination appear misguided. But this suggestion is bankrupt, since the work done by denying the Homogeneity Assumption is ensuring that mental causes and their physical competitors are treated asymmetrically. But what the empirical details outlined in this section reveal is that mental causes and their physical competitors – neurophysiological causes – should be treated on a par *vis-à-vis* the question of production: neither are connected to their effects via spatiotemporally local and contiguous processes.

consistent with the empirical details outlined above that differ significantly from (ST*). Nonetheless, we should recognize that the compatibility of the counterfactualist's principal claim with these empirical details provides some powerful reasons in favor of the counterfactualist approach. At the very least, we can conclude that the consistency of (ST*) with these empirical details makes counterfactualism a significantly better option than any rival gripped by productive causation. Therefore, I submit that the failure of Kim's arguments and the compatibility of (ST*) with our best empirical theories of muscle contraction provide us with substantial reason to adopt the counterfactualist approach to grounding mental causation.

Still, the counterfactualist's work is not complete, for there remains several issues concerning their principal claim that counterfactual dependence is sufficient for causation. To my mind, some of the least serious concerns come from Bennett (1987) and Lombard (1990). These authors present some considerations involving the delay of an event which, when combined with (ST*), generate some *prima facie* troublesome conclusions. For example, Bennett devises an example in which the heavy rains in April prevent the electrical storms in the following two months from starting a forest fire. However, in June the electrical storms persist and the forest eventually catches fire. If it had not been for the April rains, then the forest fire – the event that is the actual burning of the forest – would not have occurred (Bennett 1987, 373). The problem is that (ST*) yields the claim *that the April rains caused the forest fire* when we know that "it is a bit of good common sense that heavy

rains can put out fires, they don't start them" (Lombard 1990, 197). Menzies (2004) presents other examples purported to be problematic for (ST*). Consider the following:

> A person develops lung cancer as a result of years of smoking. It is true that if he had not smoked he would not have developed cancer. It is also true that he would not have developed lung cancer if he had not possessed lungs, or even if he had not been born (Menzies 2004, 143).

As Menzies goes on to say "it is absurd to think his possession of lungs or even his birth caused his lung cancer" (Menzies 2004, 143). The problem here is that (ST*) is insensitive to the commonsense distinction between genuine causes and mere background conditions, since it counts both the smoking and the possession of lungs (or the birth) as causes of the lung cancer.

I do not find these counterexamples persuasive because there is good reason to think that, despite the appearances, (ST*) gets the right verdict. We should take seriously a "methodological sermon" offered by Ned Hall (2004):

> If you want to make trouble for an analysis of causation – but want to do so on the cheap – then it's convenient to ignore the egalitarian character of the *analysandum*. Get your audience to do the same, and you can proceed to elicit judgments that will appear to undermine the analysis, but which are in fact irrelevant to it … (Hall 2004, 228).

So what exactly is the analysandum of (ST*)? Lewis is pretty clear that the analysis is intended to capture the "broad and nondiscriminatory concept of causation (unselectively speaking)" (Lewis 1986, 162). And the counterfactualist surely agrees that (ST*) provides a sufficient condition for being *a cause* or *among the causes* as opposed to *the cause* of some effect. Lewis tells us,

> We sometimes single out one among all the causes of some event and call it "the" cause, as if there were no others. Or we single out a few as the "causes", calling the rest mere "causal factors" or "causal conditions" … We may select the abnormal or extraordinary causes, or those under human control, or those we deem good or bad, or just those we want to talk about (Lewis 1986, 162).

These purported counterexamples confuse the "egalitarian sense of 'cause' with a much more restrictive sense (no doubt greatly infected with pragmatics) that places heavy weight on salience" (Hall 2004, 228). I submit that when the above verdicts given by (ST*) appear unacceptable it is because we have forgotten that (ST*) presents us with a condition for being *among the causes* of an effect, not what it is to be *the cause*.

So I agree with Lewis and Hall that (ST*) gets the right answers here. The April rains are *among the causes* of the forest fire in June. The possession of lungs and the birth are *among the causes* of the individual's developing lung cancer. But, in most ordinary contexts, these causes are not amongst the most salient parts of the effect's causal history and, therefore, are not what we consider to be *the cause*. Additional support for (ST*)'s verdict can be garnered from the following considerations. Take, for instance, the example of the development of cancer and the individual's possession of lungs. The possession of lungs is *statistically relevant* to developing lung cancer, since the probability of getting lung cancer given the circumstances and possessing lungs is greater than getting lung cancer given the circumstances and not possessing lungs. The possession of lungs may provide *predictive evidence* of lung cancer just as lung cancer provides *retrodictive evidence* of the possession of lungs. A *complete explanation* of why someone

develops lung cancer would surely include their possession of lungs. And, finally, the possession of lungs is an *effective way* for an agent to manipulate whether one gets lung cancer. Many of the conceptual connotations of causation are at work here and this suggests, just as (ST*) implies, that the possession of lungs is among the causes of the development of lung cancer.[92] The same considerations apply fairly straightforwardly to the other examples.

But there is a problem facing the counterfactualist that is not so easily disarmed. Consider the following well-known case of *causation by omission*:

> Jones plans on leaving town for a few weeks and asks his neighbor, Smith, to take care of the plants in his garden. Smith agrees, but neglects to care for the plants by failing to water them. The plants wilt and eventually die.

One mark in favor of (ST*) is that it easily captures the intuitively correct verdict that Smith's non-watering of Jones's plants causes them to wilt. The wilting of the plants is counterfactually dependent on Smith's omission, as revealed by the truth that the plants would not have wilted had Smith watered them. But, as opponents of (ST*) have often pointed out, the wilting of the plants is counterfactually dependent on numerous omissions that intuitively are *not* amongst the causes of the wilting. For instance, if the Queen of England had watered the plants in Jones's garden, then they would not have wilted. Therefore, given (ST*), it follows that the Queen of England's non-watering of the plants is among the causes of their wilting.

---

[92] Again, see (Schaffer 2000a, 2000b, 2004) for this kind of argumentative strategy.

The problem is more widespread than this single example illustrates. If (ST*) is correct, then there are an indefinite amount of absences and/or omissions included in the causal history of an event that we have *strong intuitions* are not included in the event's causal history. Here are some particularly striking examples. Among the causes of Jones's asking his neighbor Smith to water his plants are (i) the absence of a massive earthquake occurring in the region occupied by Jones right before the request, (ii) the absence of an aneurism in Jones's brain moments before the request, (iii) the absence of nerve gas in Jones's house a few hours before the request, and, going even further back in time, (iv) the omission by Killer to murder Jones's father as a young boy, etc. The problem is that, while (ST*) accommodates the intuitive cases of "negative causation", it allows far too many others in at the same time. A complete defense of counterfactualism demands that something be said to mitigate the profligate manner in which (ST*) introduces absences and omissions into an event's causal history. Unfortunately, I cannot pursue a resolution of this problem in the present essay.[93] Nevertheless, a strong case for counterfactualism has been made with the caveat that something needs to be done to curb (ST*)'s generous admission of "negative causation".

[93] However, see (Schaffer 2004, forthcoming), (Hitchcock 2007), (Russo and Montminy, forthcoming) for potential solutions to this problem.

Chapter 5

The Exclusion Dilemma: The Overdetermination Horn

Jerry Fodor (1989) writes,

> If it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying … if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world (Fodor 1989, 77).

Throughout the previous chapter, I argued that we can avoid this "Fodorian" apocalypse if we ground mental causation in counterfactual dependence between distinct events. At the very least, I hope to have shown that the counterfactualist remains in a much better position with regards to vindicating the difference-making causal status of the mental than any production theorist about causation.

But, in doing so, I have incurred a special burden that I hope to discharge in the present chapter. If mental events are distinct from their neurophysiological realizers *and* every bodily effect with a cause has a neurophysiological cause *and* these same bodily effects have a mental cause, then I have walked right into the problem of causal overdetermination. In this chapter, I argue that overdetermination is a problem for three reasons, but a proper analysis of the relation between mental and neurophysiological events provides the resources for undercutting these worries.

In Section 5.1, I discuss the three underlying concerns suggested in the literature as to why causal overdetermination is troublesome. If mental causation involves overdetermination, then either mental causes and their neurophysiological

realizers are causally dispensable *vis-à-vis* their overdetermined effects, mental causation involves an unexplained coincidence or conspiracy, or overdetermined effects receive a duplicative transfer of "causal oomph!" which entails a failure of a particular sort of physical explanation. In Section 5.2, I present a rough outline of how the threat from overdetermination should be handled by appealing to the mind-body relation. I discuss how a nonreductive physicalism which holds that mental events are *necessitated* by their neurophysiological realizers deals with these three concerns. Additionally, I consider a nonphysicalist mind-body relation and argue that it cannot deal with the problem of causal dispensability. Finally, in Section 5.3, I present an analysis of *event realization* and show that it is able, with some of our conclusions from Chapter 4, to overcome all of the difficulties associated with overdetermination. If my arguments are successful, we would have a vindication of mental causation without the threats associated with causal overdetermination, all of which is consistent with nonreductive physicalism.

Section 5.1: Why Overdetermination is Bad

The primary concern of this chapter is the tension between the claim that both a mental event and its neurophysiological realizer are among the causes of some bodily effect and that this kind of overdetermination is *not* a problem. As Karen Bennett (2003) remarks, "The more you go out of your way to establish the full-fledged efficacy of the mental, the more it sounds like its effects are overdetermined" (Bennett 2003, 472). Perhaps we should insist that our goal is to show that the nonreductionist picture of mental causation that has emerged from

our discussion is not *really* overdetermination at all. I think we could insist on this goal, but what we really want to show is that it is *not* problematic for a single bodily movement to be an effect of both a mental cause and its neurophysiological realizer. Whether we label this as not really overdetermination at all or an *unproblematic* form of overdetermination is a terminological issue. In the interest of clarity, I assume that the situation which presents the nonreductive physicalist is a form of overdetermination, which I shall refer to as MN-overdetermination. The goal of the present chapter, then, is to argue that "the mental/physical case is *importantly different* from the standard textbook examples of firing squads, houses that are struck by lightning at the same moment that someone tosses a lit cigarette into the draperies, and so forth" (my emphasis) (Bennett 2003, 474). What the nonreductive physicalist needs to do is "break the analogy" between cases of MN-overdetermination and standard cases of overdetermination.

At the end of the day, those who endorse the exclusion argument as a serious dilemma facing the nonreductive physicalist claim that the analogy is an apt one, that is, the differences in MN-overdetermination are not important enough to avoid the problems associated with standard cases of overdetermination. But a frequently unaddressed issue in the literature on the exclusion problem is *why* overdetermination is an outcome to be avoided. This point bears some emphasis, for it is a crucial premise in the exclusion argument that overdetermination is problematic and should not be welcomed as a consequence. A particularly obvious reply may be that MN-overdetermination is supposed to be a pervasive

phenomenon. In other words, if mental causes are as *pervasive* as we pretheoretically think they are, then the nonreductive physicalist committed to MN-overdetermination is committed to *systematic overdetermination*. And if it is systematic, it is a problem.

Oftentimes, this is the view implicit in the literature. For instance, Trenton Merricks (2001) tells us that a substance dualist who concedes that the physical world is causally complete is "pre-theoretically" committed to "an ugly picture" and that "the redundancy is *all by itself* a reason to resist this form of substance dualism" (original emphasis) (Merricks 2001, 67). More generally, he claims that "we always have reason to resist systematic causal overdetermination, along with any view that implies it" (Merricks 2001, 67). Jaegwon Kim, the most notable proponent of the exclusion problem as a reason to move back towards type-reductionism, can occasionally be read in a similar way, endorsing the view that systematic overdetermination should be avoided at all costs. He asks, "If C and C* are each a sufficient cause of the event E, then why isn't E overdetermined? It is at best extremely odd to think that each and every bit of action we perform is overdetermined in virtue of having two distinct sufficient causes" (Kim 1989, 86). What Merrick's and Kim's remarks suggest is that systematic overdetermination is just bad *punkt*.

However, it is reasonable to ask why the *frequency* in which the overdetermination occurs makes it troublesome. That systematic overdetermination is problematic is an indispensable premise in Merrick's primary

argument for the elimination of inanimate macro-physical objects and in Kim's central critique of nonreductive physicalism. It seems to me that once this premise is employed in controversial philosophical argument, it would be bad philosophy to leave it unquestioned. Certainly, it is not as if systematic overdetermination "wears" its badness on its face. I am in agreement with Ted Sider (2003) when he writes that there is a reason why overdetermination is bad and, until the source of the badness is specified, the complaint is philosophically useless (Sider 2003, 721). In the next few sections, I attempt to give some content to the idea that systematic overdetermination is problematic.

Section 5.1.1: Causal Dispensability

One very natural reason for taking overdetermination to be problematic is that it involves some kind of *causal redundancy or dispensability*.[94] To bring this idea into sharper focus, consider the standard case of overdetermination first introduced in Section 1.5.2:

> Assassin and Badgirl simultaneously poison Victim's coffee with identical doses of a lethal poison. Either dose by itself would have sufficed for Victim's death. Victim drinks the coffee and dies. He would have survived if the coffee had not been poisoned.

What we see here is that, given the circumstances, neither Assassin's nor Badgirl's poisoning of the coffee is *needed* for the occurrence of Victim's death. This is why each cause is considered to *suffice* for the effect and, furthermore, why this is not a case of *joint causation*. If we wanted to explain why Victim died on this occasion, we would only need to cite either Assassin's action or Badgirl's action; there is no

[94] Something like this is suggested in Menzies (2003) discussion of the exclusion problem.

need to cite both. For example, we could offer a fully satisfactory explanation of the death without even mentioning Assassin's poisoning of the coffee. This makes Assassin's action a causally dispensable part of the death's causal history.

In order to assess this proposal, we need to assign a more concrete sense to the claim that overdetermining causes are causally dispensable. A natural way of understanding this is as follows:

> (Causal Dispensability) When c is an overdetermining cause of e, c is *causally dispensable* with respect to an effect e just in case e would still have occurred if c had not occurred.

I have defined causal dispensability in terms of the counterfactual 'if c had not occurred then e would still have occurred' so I must specify how the counterfactual is to be evaluated. I submit that the evaluation proceeds in the manner specified by Lewis's similarity metric, that is, we hold the past fixed as much as possible, countenance an inconspicuous violation of law – a small, local miracle – resulting in the non-occurrence of c, and determine whether the effect e still occurs. At this point it should be fairly obvious that saying an overdetermining cause is causally dispensable with respect to its effect is equivalent to saying that the effect does *not* counterfactually depend on the overdetermining cause. This is exactly what we should expect: the presence of one overdetermining cause breaks the dependency between the effect and the other overdetermining cause.[95] It is straightforward to see that both Assassin's action and Badgirl's action are causally dispensable with

[95] Isn't it problematic to claim that an overdetermined effect is *not* counterfactually dependent upon either overdetermining *cause*? It is if counterfactual dependence is necessary for causation, but this is not the relation between causation and counterfactual dependence defended in the last chapter.

respect to Victim's death. Victim's death would still have occurred if either Assassin's action had not occurred or if Badgirl's action had not occurred.

According to this proposal, in that very same sense, mental events are *causally dispensable parts* of their effect's causal history. This fact alone might not be so troublesome if MN-overdetermination is infrequent. But this is precisely *not* the situation the nonreductive physicalist is committed to, since they claim that every instance of mental causation is concurrent with an instance of causation by its distinct neurophysiological realizer. The problem, then, is that the nonreductive physicalist makes mental events causally dispensable *across the board*. Every mental event is a dispensable part of its effect's causal history. Of course, we can say the same thing about the neurophysiological realizers of those mental events. Kim (1993b) insists that this has a rather unfortunate consequence. He tells us that "the overdetermination idea seems to violate the causal closure principle as well: in the counterfactual situation in which the physical cause does not occur, the closure principle is violated" because "if the physical cause hadn't occurred, the mental cause by itself would have caused the effect" (Kim 1993b, 281). Systematic overdetermination is bad because it makes both mental and neurophysiological events dispensable across the board. In addition, the causal dispensability of neurophysiological events seems to violate the "causal closure principle".

Before moving forward, we should be clear that the above proposal does *not* claim overdetermining causes *must* be causally dispensable with respect to their overdetermined effect. To see that this is not a requirement on overdetermination,

suppose that c_1 and c_2 each cause an effect e and that there is some mechanism present in the circumstances with the following features: it is set up to prevent c_2 from causing e but the presence of c_1 inhibits or blocks the mechanism from preventing c_2 in this way. Under such circumstances, it is false that if c_1 had not occurred then e would still have occurred, for if c_1 failed to occur, the mechanism would no longer be inhibited and thus would prevent c_2 from causing e.[96] All of this seems consistent with c_1 and c_2 overdetermining the effect e, which implies that causal dispensability is not necessary for overdetermination. All that is proposed is that overdetermination often involves causal dispensability so *systematic* MN-overdetermination involves a lot more of it. The result is that the nonreductive physicalist makes mental events causally dispensable to their effects and this is a rather unwelcome consequence.

Section 5.1.2: Coincidence and Conspiracy

In the literature, one is often confronted with the claim that overdetermination is problematic because it involves *unexplained correlation or coincidence*. Ted Sider (2003) writes:

> Imagine a paranoiac who thinks that every time someone is shot, there are two causally independent shooters. He is crazy, but why? One reason (not the only one) is that it would be a coincidence that all these sharpshooters just happen to fire at the same places at the same times. This great regularity would need an explanation, and none could be given. Likewise, it may be claimed, widespread overdetermination … by mental and physical causes, would require a massive, unexplained correlation between the multiple causes (Sider 2003, 722).

[96] See (Simona 2011, 475) for an example of this kind.

The suggestion seems correct. Overdetermination often involves a coincidence of some kind. An unlucky victim is shot *independently* by two assassins at the same time. The cigarette *just happens* to ignite the curtains at the exact moment the house is struck by lightning. Given the ubiquity of MN-overdetermination on the nonreductive physicalist picture, these coincidences would have to be pervasive features of our world. Such widespread and massive unexplained correlation is surely a reason why systematic overdetermination is bad.

Again, we should be careful to not misconstrue the proposal as claiming that coincidence is *necessary* for overdetermination. A simple addition to our example from Section 1.5.2 demonstrates why.

> Assassin and Badgirl simultaneously poison Victim's coffee with identical doses of a lethal poison *as a part of a carefully crafted scheme to ensure Victim's death*. Either dose by itself would have sufficed for Victim's death. Victim drinks the coffee and dies. He would have survived if the coffee had not been poisoned.

Assassin's action and Badgirl's action overdetermine Victim's death, but it is not true that this correlation lacks an explanation. This is no coincidence because Assassin and Badgirl decided to work together, concocting their plan such that each poisons the coffee at the same time. Or perhaps their actions were part of a larger conspiracy against Victim led by Cautious, who is well-known for employing two assassins to complete the same job to increase the probability of success. Whatever the exact details, it should be clear that overdetermination does not require coincidence.

This last observation suggests we extend the proposal in the following manner. Suppose our paranoiac thinks instead that every time someone is shot, there are two shooters working in tandem as part of some larger conspiracy. He may not be as crazy as our first paranoiac, but he is, at the very least, irrational. His explanation for why all these sharpshooters fire at the same places at the same times is *prima facie* implausible. If conspiracy is appealed to in order to avoid unexplained correlation, then widespread overdetermination with its widespread coincidence would seem to demand widespread conspiracy. The implausibility is compounded. Similarly, MN-overdetermination looks to require the postulation of some sort of widespread conspiracy in order to avoid coincidence. This result looks significantly worse for the nonreductive physicalist, for what, short of Leibnizian pre-establish harmony, could explain the correlations between mental and neurophysiological causes?

Section 5.1.3: Duplicative "Causal Oomph!"

Often, the lesson drawn from the exclusion argument is not that the mental is stripped of its causal powers, but rather that the causal power it has is of a different sort. More specifically, the mental possesses a *derivative* causal status, a causal status it acquires from its relationship to causes of a more fundamental variety.[97] Motivated partly to respect the "closed character of physical theory", Kim (1984) presents his account of supervenient or epiphenomenal causation: "When a mental event M causes a physical event P, this is so because M is

[97] See Section 2.2.

supervenient upon a physical event, P*, and P* causes P" (Kim 1984, 267 – 268). As he goes on to tell us,

> It would be foolish to pretend that the proposed account accords to the mental *the full causal potency* we accord to fundamental physical processes … Mental causation does take place; it is only that it is epiphenomenal causation, that is, a causal relation that is reducible to, or explainable by, the causal processes taking place at a more basic physical level (my emphasis) (Kim 1984, 268).

But we could just as easily see this account motivated to avoid the purported problems associated with overdetermination. If both a mental event and its neurophysiological realizer cause some bodily effect, then *this* is overdetermination and *that* is a problem. Well, it would only be a problem if we held that the mental is afforded full causal potency in addition to the potency possessed by the neurophysiological. Kim's account of "epiphenomenal causation" avoids the threat of overdetermination because the mental's causal potency is derived from, reducible to, or otherwise explained in terms of, the potency had by the neurophysiological.

Likewise, Jackson and Pettit (1990) are motivated by exclusion-like concerns and argue that being efficacious in the production of an effect is not the only way to be causally relevant to that effect (Jackson and Pettit 1990, 112). On their view, there are "at least two distinct ways in which a property can be causally relevant: through being efficacious in the production of whatever is in question, or through programing for the presence of an efficacious property" (Jackson and Pettit 1990, 115). Extending and applying this to mental *events*, we can say that mental events are relevant to their effects, not by being efficacious in the production of

these effects, but rather by "programing for" some efficacious event. The mental event "does not figure in the productive process leading to the [effect] but it more or less ensures that a property-instance which is required for that process does figure" (Jackson and Pettit 1990, 114). Once again, the threat posed by overdetermination is avoided, since mental events do not exercise their causal potency in the same way as their neurophysiological realizers. The mental is merely "instrumentally effective" with respect to its bodily effects, since its realization is "a good tactic for producing the effect", not by producing it *directly*, but by *ensuring* or "programing for" something that does (Jackson and Pettit 1990, 109).

I believe the project of avoiding the threat of overdetermination by claiming that the mental enjoys a "lesser grade of causal efficacy" (Levine 2001, 28) betrays the following underlying worry about overdetermination. Schaffer (2004b) remarks,

> If causation is taken to be a primitive (and at times a directly observable) relation, then it is hard to resist the idea of "causal oomph!" with its disastrous implications that overdetermination should generate excess "*oomph!*" (– the person hits twice as hard, or jumps twice as far) (Schaffer 2004b, 237 – 238).

Similarly, the thought is that *full causal potency* or *causal efficacy* easily leads one to the idea that real, genuine causation involves some "oomph!". This implies that the nonreductive physicalist's commitment to MN-overdetermination involves a duplicative transfer of causal "oomph!" to overdetermined bodily effects.

The idea of "causal oomph!" deserves to be elucidated, but at the very least it conveys a picture of causation where events *transfer* something to their effects. Schaffer (2004a) notes that the views defended in the literature that take "causal oomph!" seriously share the idea that causation requires a *physical connection* between cause and effect, since a connection is needed in order to transfer the "oomph!" (Schaffer 2004a, 203 – 204).[98] Their differences lie in what must persist through time and be transferred from cause to effect. According to Aronson (1971), Fair (1979), and Castaneda (1984), causation requires the transfer of the property of energy-momentum from cause to effect. Dowe (1992, 1995, 2000) following Salmon (1998) defends a similar view by characterizing "causal processes" as involving the persistence through time of an object possessing some conserved physical quantity and a "causal interaction" as an intersection of causal processes involving an exchange of the conserved physical quantity. Ehring (1997) argues that causation requires the transfer of a trope from cause to effect and Kistler (1998, 2001) limits these to tropes of conserved quantities.

Whatever the precise details of the story, the problem with overdetermination if "causal oomph!" is taken seriously is the failure of a certain type of physical explanation. The nonreductive physicalist committed to MN-overdetermination who admits that the mental has *full causal potency* must allow

[98] Dowe (2004) expresses nicely this idea: "Suppose I had thrown the rock through the window. Then my throwing the rock caused the window to break precisely because there is a causal process, the trajectory of the rock, possessing momentum, which *links* my throw to the window's breaking. And the window's breaking involves an *exchange* of momentum" (my emphasis) (Dowe 2004, 189). Causation requires a connection or *linkage* from cause to effect in order to transfer or *exchange* something.

that there are *two connecting processes* running from both the mental event and the neurophysiological event to the bodily effect. Kim (1984) writes,

> It is hardly conceivable that the pain sensation *qua* mental event acts directly on the muscles of my arm, causing them to contract … If the pain is to play a causal role in the withdrawal of my hand, it must do so by somehow *making use of* the usual physiological causal path to this bodily event; it looks as though the causal path from the pain to the limb motion must *merge* with the physiological path at a certain point (Kim 1984, 265 – 266).

But the concession that the two connecting processes must "merge" has an undesirable consequence:

> If there is such a point, that must be where psychophysical causal action takes place … [but] there is the deeper problem that any such nonphysical intervention in a physical system would jeopardize the closed character of physical theory. It would force us to accept a conception of the physical in which to give a causal account of, say, the motion of a physical particle, it is sometimes necessary to go outside the physical system and appeal to some nonphysical agency and invoke some irreducible psychophysical law (Kim 1984, 266).

Recall that these connecting processes involve the transfer of something from cause to effect. Call whatever precisely is transferred 'Q'. The point of "psychophysical causal action" is where the mental and neurophysiological connecting processes merge. This intersection involves *the exchange of Q*, which implies that some part of the remaining connecting process leading to the bodily effect – say, some physiological event p* in the body – receives an additional amount of Q. In other words, the event p* instantiates additional energy-momentum or includes more of the "conserved-quantity-bearing object" or has an additional conserved quantity trope. The fact that p* has some additional Q is something that cannot be explained physically. No antecedent physiological event nor any

neurophysiological event can explain why p* has as much Q as it does. Therefore, we must leave "the realm of the physical" in order to explain this fact about p*.

Kim suggests that this type of failure of physical explanation is an outright violation of the "closed character of physical theory", or the assumption of Completeness. Regardless of whether this is so, we can at the very least conclude that this type of failure in physical explanation does not sit comfortably with those who endorse Completeness. Furthermore, the systematic nature of MN-overdetermination ensures that this type of failure in physical explanation is ubiquitous in our world. This is a powerful reason to reject MN-overdetermination. When this is the worry associated with overdetermination working in the background, it becomes rather natural to avoid the threat of double "oomph!" by either holding that the mental *inherits* its "oomph!" from its realizer (e.g., Kim (1984)) or merely *ensures* an "oomphy!" realizer will be there to do the causing (e.g., Jackson and Pettit (1990)).[99]

Section 5.2: The Mind-Body Relation

Karen Bennett's (2003, 2008) treatment of the exclusion problem provides a particularly clear discussion of what the nonreductive physicalist's task is and presents some novel ideas worth repeating. For our purposes, the principal claim Bennett offers us is that the analogy between MN-overdetermination and standard cases can be broken when we realize that "there is an important tight relation

[99] I think the same issues concerning "oomphy!" causation are working in the background of Kim's causal inheritance principle. See Section 1.5.2. If having a causal power is being disposed under certain conditions to cause something in an "oomphy!" fashion, then when an instance of a functional property is said to share its causal powers with the instance of its realizer, the "oomph!" is shared and the troublesome duplication avoided.

between the mental and the physical" that does not hold between the causes in standard cases of overdetermination (Bennett 2003, 475). In other words, for some mental event m and its neurophysiological realizer n, there is some relation R such that R(n, m) *and* it is due to m being so related to n that cases of MN-overdetermination do not inherit the problems associated with overdetermination.

A few particularly salient examples of an R are worth mentioning here. Suppose that the neurophysiological event n is *among the causes* of the mental event m which is among the causes of some intentional body movement b. The neurophysiological and mental event are part of a single causal chain that eventually terminates in the intentional body movement. That is, if R = causation, then n's being among the causes of m enables us to avoid the problems associated with overdetermination. First of all, neither m nor n are causally dispensable with respect to b. If the neurophysiological event n had not occurred, then neither would have the mental event and thus the bodily effect would have not occurred. If the mental event would not have occurred, the neurophysiological event still would have, but it would have failed to cause the bodily effect, since the causal chain leading to b would have been broken. Second, the correlation between the causes of the bodily effect is not unexplained nor is there any need to appeal to some sort of conspiracy. Both events are among the causes of the bodily effect because these events are part of a single causal chain leading to it. Finally, the concerns with duplicating "causal oomph!" are misplaced. If there is a connecting process running from the event n to the effect b, then the event m "makes use" of that very

process, since it is itself an effect of n. Hence, whatever is transferred from m to b via this connecting process is something that was previously transferred from n to m. There is no threat of failing to physically explain why the bodily effect has as much of the transferred quantity as it does, since an appeal to the neurophysiological effect suffices here. In fact, if R = causation, then we are disinclined to even consider the situation to involve overdetermination at all so *a fortiori* it does not involve a problematic form of overdetermination.

Of course, the nonreductive physicalist does not claim that the relation between mental events and their realizers is causation. But consider what *some* nonreductive physicalists do say, namely R = token-identity.[100] If the mental event is identical to the neurophysiological event, then all the problems associated with overdetermination are solved. Neither event is causally dispensable to the bodily effect. If the mental event had not occurred, then neither would have the neurophysiological event (given that m = n) and so the bodily effect b would have failed to occur. Similarly, if the neurophysiological event had not occurred, the mental event would not have either (given that n = m) and so the bodily effect would not have occurred. The correlation between the causes is not unexplained nor is there a need to appeal to a conspiracy, since the events are one and the same. In fact, as J.C.C. Smart writes in his discussion of type-identity, speaking of type-identical properties as being *correlated* carries the implication that they are "something 'over and above'. You cannot correlate something with itself" (Smart

[100] See, for instance, (Davidson 1980d).

1959, 142). Likewise, according to the token-identity theorist, saying there is some unexplained correlation between m and n is to convey something false about the situation. Similarly, to say that the event m "makes use" of the connecting process running from n to b carries the false implication that m is distinct from n. But under the present assumption this is false and so whatever connecting process runs from n to b is the very same process that runs from m to b. The event identity entails that only a single "dose" of the transferred quantity makes its way from m and n to b. We can physically explain why b has as much of the quantity as it does. Finally, if R = token-identity, then the situation does not involve overdetermination and so *a fortiori* does not involve a problematic kind of overdetermination.[101]

Now, it should have been obvious from the start that R = token-identity provides a simple and straightforward solution to the problem of overdetermination. But there are nonreductive physicalists, like myself, who take there to be good reason to reject the token-identity of mental and neurophysiological events. In Section 1.2.2, we saw some modal arguments against token-identity that appealed to a difference in *de re* modal properties to conclude that mental events are distinct from neurophysiological events. To my mind, these arguments have much *prima face* plausibility, but whether or not they ultimately can be made to work is a complex philosophical question in its own right. Henceforth, what I shall be

[101] This is why a token-identity theory offers an immediate dissolution of the overdetermination horn of the dilemma. But, as mentioned in Sections 1.4 and 2.4, most discussions of this position in the literature quickly shift focus to the epiphenomenalist horn, mental *properties*, and whether these are at all *relevant* to the causal relations entered into by mental events. However, Davidson's nominalist ontology enables him to dissolve these type-epiphenomenalist worries. Mental events are mental "only as described" (Davidson 1980d, 215) and not in virtue of instantiating some mental properties. There are no mental properties and so the concern that they may end up being irrelevant to causation is misplaced.

primarily concerned with in this chapter is what a nonreductive physicalist *who rejects token-identity* can say in response to the overdetermination horn of the exclusion dilemma. My project, then, is to find some *non-identity* relation R that holds between mental and neurophysiological events that provides similar explanations for why MN-overdetermination does not inherit the problems associated with typical cases of overdetermination.

Section 5.2.1: The Determinate-Determinable Relation

Some immediate candidates for a non-identity R that may help avoid the problems associated with overdetermination are the various forms of *asymmetric necessitation* discussed in the literature, which hold that mental events are only weakly modally distinct from their neurophysiological realizers.[102] As an example of such a mind-body relation, consider Yablo's (1992b) claim that the mental events are *determinables* of their neurophysiological realizers, that is, R = determination. As we discussed in Section 3.2.1, Yablo argues that mental *properties* are determinables of the neurophysiological *properties* that realize them. As a reminder, intuitively, we can capture the idea as follows:

> (Property Determination) Where P and Q are properties, P determines Q just in case for a thing to be P is for it to be Q, not *simpliciter*, but in a specific way (Yablo 1992b, 252).

The determinate-determinable relation is the "paradigm of one-way necessitation" (Yablo 1992b, 250), which means that P determines Q only if (a) necessarily, for

[102] I am here considering asymmetric necessitation to pick a one-way *metaphysical* necessitation relation and not some restricted form of necessitation (e.g., nomological). See Section 1.2, fn. 6. Metaphysical necessitation is, of course, the kind of necessitation involved in Yablo's understanding of determination and Shoemaker's definition of realization.

all x, if x has P then x has Q; and (b) possibly, for some x, x has Q but lacks P.[103] Things are a bit trickier if we want to explicate a determination relation between *events*, but a similar intuitive characterization should suffice for present purposes.[104]

> (Event Determination) Where p and q are events, p determines q just in case for p to occur (in a possible world) is for q to occur (there), not *simpliciter*, but in a certain way (Yablo 1992b, 260).

The bolt's sudden snapping determines the bolt's snapping *per se* because for the bolt to suddenly snap in this world is for the bolt to snap in this world, not *simpliciter*, but in a certain way (viz., suddenly). And again, it should be understood that event determination is an *asymmetric necessitation* relation. You cannot have the bolt suddenly snap without it snapping *per se* though you can have it snap *per se* without it snapping suddenly.

Let us assume, then, that R = determination and that R(n, m). The neurophysiological realizer n of some mental event m determines that mental event such that for m to occur in the actual world is for n to occur there, not *simpliciter*, but in a certain way. Additionally, necessarily: if n occurs, m occurs but possibly:

---

[103] Additionally, Yablo tells us there is no presumption that the necessitation of Q by P is *a priori* knowable. Perhaps it is (e.g., the property of being red necessitates the property of being colored) or perhaps it is not (e.g., the property of being K, some highly specific micromechanical state, necessitates the property of being at temperature 95 degrees C) (Yablo 1992b, 252 – 253).

[104] The problem, as Yablo states, is that "determination involves the idea that the requirements associated with one thing include the requirements associated with another; and although properties are requiremental on their face, particulars are not" (Yablo 1992b, 261). His solution is to appeal to the notion of *individual essence*. The idea is that an event p determines an event q only if: (a) necessarily, if p exists, q exists and is "categorically indiscernible" from p; and (b) possibly, q exists and p does not. Roughly, an event p is categorically indiscernible from an event q when p's essence includes all of q's essence *plus a bit more* (i.e., p subsumes q) and, in the world of occurrence w, the way p is in w is just like the way q is in w *with respect to their categorical properties*. The obvious circularity here should not matter for our purposes. See (Yablo 1992b, 263 – 265) for the details.

m occurs without n occurring. This entails a commitment to what, in Section 1.3.1, I labeled L-Nonreductionism, for mental events *locally supervene* on neurophysiological events:

> (M-N Local Supervenience) Any minimal neurophysiological duplicate of an actual individual is a psychological duplicate *simpliciter* of that actual individual.

Where N_i reports all the neurophysiological facts about a specific individual and M_i reports all of that individual's mental facts, if neurophysiological events determine mental events, then 'if N_i then M_i' is a *necessary truth*. In short, neurophysiological duplicates are psychological duplicates.[105] What is most relevant for present purposes is how we can appeal to R = determination to "break the analogy" between MN-overdetermination and typical cases of overdetermination.[106]

Immediately, we can see that the coincidence and conspiracy concern simply evaporates. Consider the standard case of the determinate-determinable relation mentioned previously. Every time the bolt suddenly snaps, it snaps *per se*. These events are correlated and they are *systematically* correlated, but the correlation is *not* unexplained. The snapping of the bolt is determined by its suddenly snapping. Similarly, the systematic correlation between mental events and their neurophysiological realizers can be explained by the fact that neurophysiological events determine mental events. As Block (1990) has pointed

[105] We might wonder whether this commits Yablo to a rejection of anti-individualism. This will be addressed at the end of this section.

[106] I am examining Yablo's position without a commitment to the proportionality thesis; thus, causes have to be neither required or enough for their effects in the sense defined in Section 3.2.2.

out, "We are normally reluctant to accept overdetermination because it is wrong, other things equal, to postulate coincidence" (Block 1990, 159), but no such coincidence would be involved when R = determination. Furthermore, the correlation between the mental and neurophysiological causes is not the result of an implausible conspiracy or a Leibnizian pre-established harmony. Rather it is the result of the asymmetric necessitation relation that holds between mental events and their neurophysiological realizers. If R = determination, then MN-overdetermination is quite unlike standard cases of overdetermination which regularly involve coincidence or conspiracy.

The problems concerning causal dispensability are also avoided by appealing to R = determination. Recall in the case of R = token-identity, if the mental event had not occurred, then the neurophysiological event would not have occurred either. Our evaluation of the counterfactual relevant to determining whether m is a causally dispensable part of the bodily effect's causal history requires we consider *the most similar* not-m world. This is a world where the causal history of m is held fixed as much as possible and a small, inconspicuous violation of law results in m's non-occurrence. If the event m is identical to the event n, then it is *impossible* for the small, inconspicuous violation of law to affect the occurrence of the mental event m without affecting the occurrence of the neurophysiological event n identical with m. Hence, the most similar not-m world is also a not-n world. Thus, the bodily effect b would not have occurred and m is not a causally dispensable part of b's causal history. Similarly, if R =

determination, then if the mental event m had not occurred, the neurophysiological event n would not have occurred either. However, this is not because m and n are identical, but rather because n *determines* m. Consider the most similar world in which the bolt's snapping *per se* does not occur. This is also a world in which the bolt's sudden snapping does not occur, for it is *impossible* for the bolt to suddenly snap without snapping *per se*. Hence, if n determines m, it is impossible for n to occur without m occurring. Thus, the not-m world is a not-n world where the bodily effect b also fails to occur. If R = determination, the mental event is not a causally dispensable part of the bodily effect's causal history.

We come to the same conclusion concerning the neurophysiological event so long as we are careful to evaluate the counterfactual correctly. It is a natural thought that the most similar not-n world is a world where the mental event m still occurs. The reason is the small, inconspicuous violation of law would result in n's non-occurrence, but this non-occurrence of n would yield a very similar neurophysiological event n* to take n's place. This different though similar neurophysiological event n* is another way to realize the mental event m (remember: the mental event m is *multiply realizable*) so m's occurrence remains fixed. In a world where n's non-occurrence is replaced by n*, n* would determine m and the bodily effect b would still occur. Hence, it looks as if the neurophysiological event n is a causally dispensable part of b's causal history. However, as we saw in Section 2.4., the most similar not-n world is *not* a world where n is *replaced* by n* which also realizes m. These kinds of "replacement

readings" of counterfactuals are inappropriate in causal contexts, since we lose the counterfactual dependencies that intuitively should hold.[107] Again, it is worthwhile to cite Bennett on this point,

> When you are supposed to imagine c_1 gone, you imagine it gone. You do not worry about how the past would have to be different to make it fail to occur, and you do not worry about what else might occur in its place. You simply snip it away as though you had a metaphysical hole-puncher (Bennett 2003, 482).

So, if we make sure to avoid a replacement reading, then the world we should consider is a not-n world where n is not replaced by n*, but is simply "snipped away". In such a world, there is nothing there to realize and so determine the mental event m so the bodily effect b fails to occur. We arrive at the same conclusion: the neurophysiological event n is not a causally dispensable part of the bodily effect's causal history. This once again reveals how different MN-overdetermination is from typical cases of overdetermination if R = determination.

It is worth reiterating these points in a different way. In standard cases of overdetermination, the overdetermining causes are causally dispensable parts of the effect's causal history because, so to speak, the small, inconspicuous violation of law is able to "surgically intervene" on the events. In other words, the violation of law can "act on" one of the causes *without affecting the occurrence of the other*; it

[107] Perhaps another example will help. Suppose I seek some chocolate cake because I desire to eat some chocolate cake. Presumably, there is a *ceteris paribus* psychological law connecting those types of desires with that type of behavior. Hence, my seeking behavior is counterfactually dependent on my desire for chocolate cake. But if the most similar world where my desire for chocolate cake does not occur is a world where it is *replaced* by a different though similar desire, say a desire to eat something sweet, then my seeking some chocolate cake behavior might still occur. It follows from this that my chocolate seeking behavior is not counterfactually dependent on my desire to eat chocolate cake. Just as we must put on ban on backtracking in causal contexts, we must ban replacement readings as well.

can "differentiate" between them. And by doing so it can reveal that the effect is counterfactually dependent on neither overdetermining cause and hence why each is causally dispensable with respect to the overdetermined effect. The work done by the determination relation in this case is to bring the mental and neurophysiological events close enough together that a surgical intervention on one is not possible. Of course, it is metaphysically possible to separate the events in one direction, since determination is only an *asymmetric* necessitation relation. But, in a restricted sense of 'possible', it is not possible to separate them. At least it is not possible to separate them in the worlds that qualify as *most similar* to the actual world in causal contexts. Therefore, the small, inconspicuous violation of law cannot affect the occurrence of either the mental or neurophysiological event without affecting the occurrence of the other.

Woodward (2011) comes to a similar conclusion working in a manipulationist framework for understanding causal relations. The underlying idea of the manipulability theory of causation is that causal relations are relations that are *potentially exploitable for purposes of manipulation and control*. Roughly, events are represented in terms of variables taking certain values, so what it means for an event x to be among the causes of an event y is that the variable Y would change in value under *some suitable intervention* that changes the value of the variable X. The key notion is that of a suitable intervention which can be intuitively grasped as follows: an intervention on X with respect to Y is an exogenous causal process that changes X in such a way that if any change occurs in

Y, it occurs only in virtue of Y's relationship to X and not in any other way (Woodward 2003, 47).[108] With this in mind, Woodward (2011) argues that interventions on variables related by non-causal relationships of dependency, like determination, are exogenous causal processes that act on or change the value of *both* variables.[109] In assessing the efficacy of events related by determination, it is not appropriate to consider what would happen to the effect if the determinable is intervened on and determinate held fixed or vice versa. Similarly, in our assessments of causal dispensability, we must treat the small, inconspicuous violation of law – the Lewisian analog of an intervention – as acting on both the determinable and its determinate and discount the possibility of acting on either separately.

In short, what we see is that both the Lewisian framework endorsed in this essay and Woodward's manipulationist treatment of causation treat events that are tightly related via determination as a "unity", that is, *as if they were one and the same event* across the worlds deemed most similar to the actual world. When neurophysiological events determine mental events, a small, inconspicuous violation of law in the most similar worlds cannot surgically intervene on the events. In effect, it acts as if the events were identical. This is how the mental and the neurophysiological avoid being causally dispensable with respect to the overdetermined effect. Equivalently, if these events are treated as if they were one and the same across the most similar worlds, the overdetermined effect

[108] See (Woodward 2003, Ch. 3) for a more precise definition of intervention.

[109] See (Woodward 2011, Sections 6, 7, and 8) for the details of how to implement this requirement into a manipulationist framework and the rationale for it.

counterfactually depends on both. In the most similar worlds, if a small, inconspicuous violation of law results in m's (n's) non-occurrence, then n (m) also fails to occur and the bodily effect b does not occur. In Chapter 4, I argued that counterfactual dependence between distinct events is sufficient for causation, so accepting this entails that both the mental and neurophysiological events are among the causes of the bodily effect.

Nevertheless, there remains the problem of duplicative "causal oomph!" and the threat that such a duplication leads to a failure of a certain type of physical explanation. Unlike both R = causation and R = token-identity, it is not immediately obvious how Yablo's account of the relationship between mental and neurophysiological events avoids this worry. After all, the events are still weakly modally distinct and each is a cause of the bodily effect b. Some remarks by Yablo suggest a familiar way to handle this problem. He writes,

> One could try to counter this impression by enlarging on what has already been said, viz. that to be in pain is part of what it is to be in such and such a brain state. When one state is included in another, any influence that the first has on subsequent events is included in the influence had by the second. Brain state and pain thus *share power* in a more literal sense than the one intended: not by dividing it up between themselves, in the way that books share space on a shelf with other books, but by possessing it in common, in the way that an encyclopedia share shelf space with the volumes making it up (my emphasis) (Yablo 1997, 257).

In Section 3.2.1, we saw that Shoemaker's account of realization – an asymmetric necessitation relation – similarly involves the *sharing* of causal powers between what is realized and what does the realizing. Following Shoemaker, Jessica Wilson

(2009) develops a general account of determination appealing to the idea that determinables share in the causal powers of their determinates:

> (Powers-Based Determination) A property P determines a property Q just in case the set of powers associated with Q is a *proper subset* of the set of powers associated with P.[110]

On the assumption that causal powers are dispositions of some kind, it should be straightforward how to extend this proposal to events. On this account, the increased specificity at the heart of determination is understood as the determinate event being associated with a more specific set of causal powers than its determinable (Wilson 2009, 166).

If neurophysiological events determine mental events and this is understood in accordance with Powers-Based Determination, then it follows that the causal powers associated with mental events are a proper subset of the causal powers associated with their neurophysiological determinates. Suppose that having a causal power is being disposed under certain conditions to cause something in an "oomphy!" way.[111] If some causal power P_m of mental event m manifests, then some causal power P_n of neurophysiological event n manifests such that $P_m = P_n$. Thus, there will be only a single connecting process running from both m and n to the bodily effect b. With only a single connecting process, there will not be a duplicative transference of some quantity to the overdetermined effect b and the

[110] Wilson is well aware of the problems disjunctive and conjunctive properties pose for Powers-Based Determination. She formulates Powers-Based Determination to accommodate these problems by stating that the powers associated with P but not with Q cannot be powers associated with any property (Wilson 2009, 166).

[111] This is not to claim that either Yablo, Shoemaker, or Wilson endorse an "oomphy!" view of causation.

threat of failing to physically explain why b has as much of the quantity as it does is avoided. In summary, R = determination with the further assumption that mental events *share* their causal powers with the neurophysiological events that determine them ensures the duplicative "causal oomph!" problem is circumvented.[112]

There is no denying a striking similarity between this response to the duplicative "causal oomph!" worry and Kim's (1984) account of supervenient causation. On both accounts, mental events "make use of" the productive causal processes taking place at the neurophysiological level. However, it seems there is a difference here worth pointing out. On Kim's account of supervenient causation, mental events cause some bodily effect b *in virtue of* supervening on a neurophysiological event that causes b. In other words, the mind-body relation (i.e., supervenience) appealed to here is what affords the mental a causal status. The reverse seems to be the case with Powers-Based Determination. A mental event m is determined by a neurophysiological event n *because* the mental shares its causal powers with the neurophysiological. On this account, the causal powers in question do not "belong" more to the n than to m, nor do they belong more to m than to n. Instead, both events have an equal claim to the causal process culminating in the bodily effect and it is *in virtue of* this equal claim that the mental is determined by the neurophysiological. By reversing the direction of explanation, Powers-Based Determination does not afford the mental a causal status because of

[112] The similarities between Yablo's account of determination understood in terms of Wilson's Powers-Based Determination and Shoemaker's account of realization are apparent. So I think it is safe to say that if Shoemaker is concerned with the duplicative "causal oomph!" problem, then he could avoid it in the same way.

its relation to the neurophysiological; the mental's relation to the neurophysiological is explained by their sharing in causal powers. This avoids the concern that the mental has only a derivative causal status, a status it inherits from its relationship to the neurophysiological.

As we have seen, R = determination needs some additional assumptions about the mental's relationship to the neurophysiological in order to avoid the problem of duplicative "causal oomph!". Another option – the option I prefer – is simply to deny that the "causal oomph!" problem is really a problem at all. In short, if one rejects the underlying idea that causation requires a physical connection involving a transfer of something from cause to effect, then the worry that a certain type of physical explanation fails evaporates. Presently, I shall postpone a discussion of this option, since I take it up in later sections.

We have seen that R = determination fares quite well with respect to "breaking the analogy" between MN-overdetermination and standard cases of overdetermination. With some additional assumptions, this is a mind-body relation *consistent with a nonreductive physicalism* that can avoid the problems associated with overdetermination discussed in Section 5.1. However, a worry that remains is whether it is true that mental events are *determinables* of the neurophysiological events that realize them.[113] More generally, one might wonder how plausible it is to construe the neurophysiological as *necessitating* the mental. In Section 1.3.1, I pointed out that such a supervenience thesis, which I called M-N Local

[113] See Funkhouser (2006) for criticisms that mental phenomena are not determinables of their physical realizers.

Supervenience, is far stronger than what is required by physicalism. In short, where P reports all the physical facts and M reports all the mental facts, if physicalism is true, then 'if P then M' is a necessary truth. But it is not entailed by this that 'if N_i then M_i' is a necessary truth. The physicalist can consistently deny that neurophysiological duplicates are psychological duplicates even if they hold that any physical duplicate of the actual world is a psychological duplicate of the actual world.

Now, Yablo's position that the neurophysiological metaphysically necessitates the mental is not in itself a problem, since M-N Local Supervenience is consistent with physicalism. However, there are reasons why physicalism is typically formulated in terms of a global rather than local supervenience thesis and these reasons would speak against the view that R is some kind of necessitation relation, like determination. For instance, according to the anti-individualist, the pattern of neural firings in my prefrontal cortex could occur in a world where I have been causally-historically embedded in a different physical and/or socio-linguistic environment. Thus, this event could occur without entering into the "broad" causal-historical relations definitive of my thought that water is thirst-quenching. Given anti-individualism, this pattern of neural firings in my prefrontal cortex does not determine my thought that water is thirst-quenching; only this neurophysiological event *embedded in the right physical/social environment* determines my thought that water is thirst-quenching.

One possible response is to restrict the thesis that R = determination to just those mental events with *narrow content* and/or those mental events that are *non-intentional*. However, I do not find either option terribly plausible, since I doubt there are any non-intentional mental phenomena or mental phenomena with only narrow content.[114] Moreover, this sort of restriction leaves our task woefully incomplete. The problem of mental causation is a problem threatening the causal efficacy of *the whole range of mental phenomena* with specific concern for the propositional attitudes. It is the efficacy of my beliefs, my desires, and my intentions that sustain my capacity as an agent and plausibly anti-individualism is true of these contentful of mental events. So, if these sorts of mental events are *not* necessitated by the neurophysiological, then the problems associated with overdetermination remain and another R-relation will have to be appealed to in order to "break the analogy".

Another response is simply to reject anti-individualism about mental content. Again, this is not a position I find plausible especially for the propositional attitudes, but I will not offer any arguments for this here. Nevertheless, even the individualist about mental content should reject the thesis that R = determination. The neurophysiological events said to be the primary causal competitors of mental events are "localized brain event[s], capable of occurring in isolation from anything like [their] actual neural context" (Yablo 1992b, 270 – 271, fn. 51). Suppose that the pattern of neural firings in my

[114] Crane (2001) argues for the view that there are no purely qualitative, non-intentional mental phenomena. Tye (1995, 2000) and Dretske (1996) defend an anti-individualism about the content of phenomenal mental phenomena.

prefrontal cortex could occur outside of its actual neural context isolated in some brain matter "afloat in agar jelly". If this is a genuine possibility, it could occur in a world where it is not disposed to enter into the "narrow" causal-functional relations definitive of my thought that water is thirst-quenching. Hence, in such worlds this pattern of neural firings occurs in the absence of my thought that water is wet. Therefore, given individualism, this pattern of neural firings in my prefrontal cortex does not determine my thought that water is thirst-quenching; only this neurophysiological event *causally embedded in an appropriate neural environment* determines my thought that water is thirst-quenching.

Aware of these issues, Yablo suggests that "most mental events … seem not to be localizable in any specific portion of the brain", so we should understand mental events to be realized and thus determined by one's "overall neural condition" (Yablo 1992b, 271, fn. 51). This is problematic for two reasons. First, all along we have thought that the primary causal competitors of mental events are localized brain events and *not* an individual's overall neural condition. Suppose I am cooking some bacon and some hot oil splashes onto my hand, triggering some physiological processes terminating in a specific pattern of neural firings in my somatosensory cortex (i.e., philosophical code name: C-fiber firings). This *localized* pattern of neural firings realizes my sensation of pain. Both the neurophysiological event and the pain are amongst the causes of the retraction of my hand from the location of hot oil. What we have been primarily concerned with is the threat that these *localized* neurophysiological events pose to the causal status

of mental events. At best, Yablo's suggestion looks *ad hoc*, changing our subject to avoid the aforementioned problems. Second, we think of the realizers of distinct mental events as themselves being distinct. But if my overall neural condition realizes the pain I feel, then it also realizes every other mental event occurring in me during that time. I experience a burning sensation in my hand at the same time I experience the distinct odor of bacon cooking on a skillet and the acute hunger sensation in my stomach. Yablo's suggestion implies that all of these mental events have the same neurophysiological realizer, viz., the overall neural condition I am in during that time. This is in marked contrast to the way philosophers have originally thought of relation between mental and neurophysiological events and how we have been thinking of the exclusion problem.

What I hope to have shown in this section is that R = determination can do quite a bit of work in demonstrating the differences between MN-overdetermination and standard cases of overdetermination. But it is simply too strong of a relation to hold between mental events and "localized brain events" and it is these sorts of neurophysiological events that have been taken to threaten the causal status of the mental. In the rest of this chapter, I plan to show that there is a weaker relation, one consistent with either individualism or anti-individualism and which does not entail M-N Local Supervenience, that is capable of avoiding the concerns with causal overdetermination.

Section 5.2.2: Irreducible Psychophysical Laws

In this section, I want to consider a *nonphysicalist* mind-body relation to determine if it has the resources for avoiding the problems associated with overdetermination. Recall that the nonphysicalist rejects the truth of Global Supervenience and holds that mental events are *strongly modally distinct* from neurophysiological events. In other words, if nonphysicalism is true, then 'if P then M' is a contingent truth and, furthermore, 'if N_i then M_i' is a contingent truth. Bennett describes this general position, often called "property dualism", as such,

> The property dualist does not propose to ignore the evidence from neuroscience. He does not think that phenomenal properties float utterly free of physical properties; he thinks they are connected to physical properties in important ways. Crucially, though, he thinks the connections are merely contingent. They are on a par with the laws of science, not those of logic or metaphysics … Phenomenal properties emerge from their physical bases in some causal or quasi-causal fashion (Bennett 2005, 1).

Before we begin, it is important to note that nonphysicalists *tend* to concern themselves exclusively with phenomenal properties and events. They claim that facts about intentional mental phenomena are metaphysically necessitated by the physical facts, but facts about consciousness are not. In this section, I shall speak about mental phenomena and not make a distinction between the phenomenal and the intentional. Thus, the nonphysicalism I am concerned with thinks that neither the phenomenal nor the intentional is necessitated by the overall physical state of the world and so hold that 'if N_i then M_i' is a contingent truth, where M_i reports both the intentional and phenomenal mental facts concerning an individual.

Property dualism comes in at least three different varieties. The first is called *epiphenomenalism*.[115] According to the epiphenomenalist, mental events are contingently connected to neurophysiological events in virtue of being *the causal effects* of these events. Furthermore, mental events themselves have no physical effects. The bodily effects that appear to be caused by the mental events we undergo are caused only by the neurophysiological events which cause those mental events. In short, mental events and bodily effects share a *common neurophysiological cause*. This variety of property dualism does not face the horn of overdetermination because it accepts the horn of epiphenomenalism. We shall ignore this type of property dualism here.

The second variety of property dualism is called *emergentism*. Like the epiphenomenalist, the emergentist holds that mental events are contingently connected to neurophysiological events, but this connection is *non-causal*. Mental events are taken to "naturally supervene" on neurophysiological events in the sense that the neurophysiological facts about an individual necessitate the mental facts about that individual across those worlds that "conform to the laws of nature of our world" (Chalmers 1996, 37). The principal idea is that natural supervenience arises when "two sets of properties are systematically and perfectly correlated in the natural world … it is just a fact about *nature* that there is this correlation" (Chalmers 1996, 36). This last component is crucial, for central to the emergentist's position is that the correlation between mental and

[115] See (Jackson 1982, 133 – 136) for a contemporary defense of such a view as it pertains to "qualia".

neurophysiological events is "grounded in brute and fundamental physical-mental law-like connections (primitive 'laws of emergence')" (Kim 2006, 556). Consider the following example,

> The pressure exerted by one mole of gas systematically depends on its temperature and volume according to the law pV = KT, where K is a constant … In the actual world, whenever there is a mole of gas at a given temperature and volume, its pressure will be determined: it is empirically impossible that two distinct moles of gas could have the same temperature and volume, but different pressure (Chalmers 1996, 36).

The pressure of a specific mole of gas naturally supervenes on its temperature and volume. When the temperature and volume of a mole of gas are fixed, as a matter of law the pressure of that mole of gas is fixed. This is a contingent truth about our world, and its truth is grounded in the law pV = KT.

However, the law that holds between temperature and volume of a mole of gas and its pressure is explainable in terms of the energy-momentum of the gas molecules, the collisions between these molecules and the gas's container, and the laws describing the evolution of these micromechanical states. There is a deeper, more fundamental reason why temperature, volume, and pressure of a mole of gas are lawfully connected in the way they are. These lawful connections are *not* basic truths about our world, but explained in terms of and entailed by lawful connections between lower-level entities. But the emergentist contends that these sorts of "reductive" explanations are not available for the *non-causal* lawful connections between mental events and their underlying neurophysiological bases. These psychophysical laws are irreducible and basic truths about our world.

In addition to the irreducibility of the contingent connection between mental and neurophysiological events, the emergentist also holds that mental events have *irreducible and novel causal powers*. Horgan (2002) writes,

> [Emergentists] maintained that at various junctures in the course of evolution, complex physical entities came into being that had certain non-physical, 'emergent', properties. These properties … are fundamental force-generating properties, over and above the force-generating properties of physics; when such a property is instantiated by an individual, the *total* causal forces operative within the individual are a combination of physical and non-physical forces, and the resulting behavior of the individual is different from what it would have been had the emergent force(s) not been operative alongside the lower-level forces (Horgan 2002, 151).

What this passage suggests is that, like the epiphenomenalist, the emergentist does not face the horn of overdetermination, but for a different reason. Mental events have novel causal powers and thus have effects that do *not* have any physical causes. In short, the emergentist escapes the problem of overdetermination by rejecting Completeness.

The final variety of property dualism, the sort I want to focus on in this section, shares some features with emergentism. They hold that the connection between mental and neurophysiological events is a *lawful* connection. Furthermore, like the emergentist, they maintain that these contingent connections are irreducible and basic truths about our world. According to this sort of property dualist, we must recognize that an adequate theory of mind must posit "new fundamental properties *and laws*" (my emphasis) (Chalmers 1996, 126). However, unlike the epiphenomenalist, these irreducible psychophysical laws are *not* causal laws describing how mental events causally depend on neurophysiological events

occurring in the brain. Instead, these are "supervenience laws" informing us as to how mental phenomena "arise from physical processes" (Chalmers 1996, 127). As Loewer (2001) remarks, if our world contains such laws, then "God, when he made the world, had to make them in addition to the physical facts and laws" (Loewer 2001, 49). Additionally, this sort of property dualism is unlike the emergentist, since mental events have novel causal powers only in a weaker sense. Mental events do have physical effects, but these effects also have neurophysiological causes. Therefore, this variety of property dualism faces the same problem of causal overdetermination that faces the nonreductive physicalist. Presently, the question I want to consider is whether an *irreducible psychophysical law* connecting mental and neurophysiological events is sufficient to avoid the problems associated with overdetermination.

Recall that R = determination avoided the coincidence and conspiracy problem because the correlations between mental events and their neurophysiological realizers had an explanation. Even though the mind-body relation we are considering here is a much weaker relation than determination, the correlation still has an explanation. Mental and neurophysiological events are *lawfully connected* and so the multiple causes of an overdetermined bodily effect are not coincidentally related. The systematic correlation between mental and neurophysiological events is explained by the fundamental psychophysical laws linking certain neurophysiological types with certain mental types. Similarly, our property dualist does not have to appeal to some implausible conspiracy or

Leibnizian pre-established harmony in order to explain the correlation between the overdetermining causes. Therefore, if the mind-body relation is an irreducible psychophysical law, then MN-overdetermination does not involve coincidence or conspiracy like standard cases of overdetermination.

Now consider the problem of duplicative "causal oomph!" and the threat that a certain kind of physical explanation fails. If our property dualist admits that causation is a *productive relation* involving connecting processes, transference of quantities, and the like, then they will be hard pressed to deny that MN-overdetermination entails a systematic failure of a certain sort of physical explanation. If mental events transfer some "oomph!" to their effects not transferred by the neural, then MN-overdetermination is a problem. But, I do not think that our property dualist needs to be committed to productive causation. The duplicative "causal oomph!" problem is only a problem if causation is taken to involve a connecting process running between cause and effect. Nothing in our dualist's position entails this. If one is inclined towards a view that eschews this productive understanding of causation, even the sort of property dualism we are considering here can avoid this worry about causal overdetermination.

However, the problem of causal dispensability is not so easily avoided. Central to our property dualist's position is the claim that 'if N_i then M_i' is a contingent truth grounded in irreducible *psychophysical laws* connecting the mental with the neurophysiological. When a fundamental psychophysical law connects the mental event m with its neurophysiological base n, the most similar not-m world *is*

a world where the bodily effect b still occurs. In order to see this, recall that Lewis's similarity metric informs us that similarity amongst worlds ought to be measured with respect to two features: region of match of particular fact and conformity to laws. Furthermore, matching of particular matters of fact is *more important* than conformity to laws, so a small, inconspicuous violation of law should be tolerated at the expense of maximizing the region of match of particular matters of fact. Therefore, the most similar world to the actual world is one where the history of the actual world is held fixed as much as possible, but an inconspicuous violation of law results in the non-occurrence of the antecedent event. Now, *ex hypothesi*, an irreducible psychophysical law connects the antecedent event m with its neurophysiological base n. Hence, a counterfactual world where this psychophysical law is violated is a world that perfectly matches the history of the hypothesized actual world at the expense of a single, inconspicuous violation of law. In this counterfactual world, the mental event m fails to occur, n still occurs, and so the bodily effect still occurs. Therefore, if mental events are linked to neurophysiological events via *irreducible psychophysical laws*, they are causally dispensable parts of their bodily effect's causal histories.[116]

The neurophysiological events that serve as the physical bases of the mental events in question do not suffer the same fate. If the neurophysiological event n

[116] Loewer (2001) comes to pretty much the same conclusion when he argues that bodily effects do not counterfactually depend on mental events linked by irreducible psychophysical law to neurophysiological events. See Loewer (2001, 51 – 52). Furthermore, there is no obstacle to applying this same line of reasoning to those property dualists that insist these psychophysical laws are *causal*.

had failed to occur, then the bodily effect b would not have occurred either. The most similar world to the hypothesized actual world involves an inconspicuous violation of law that "breaks" a causal link between n and one of its proximate causal antecedents and does *not* replace n with any distinct but similar neurophysiological event n* that is also linked to m by law. In other words, as long as we are careful to avoid so-called "replacement readings", the most similar world where n fails to occur, m fails to occur, and so the bodily effect b does not occur. Hence, the neurophysiological event n is *not* a causally dispensable part of the bodily effect's causal history even if it is linked by an irreducible psychophysical law to the mental event m.

In conclusion, if the mind-body relation is an irreducible psychophysical law, cases of MN-overdetermination differ from standard cases in that *one cause but not the other* is causally dispensable with respect to the overdetermined effect. This asymmetry might be thought to be enough to "break the analogy" with standard cases of overdetermination, but it nevertheless remains the case that mental events are causally dispensable with respect to their bodily effects. The systematic nature of MN-overdetermination entails that mental events are systematically redundant and ultimately unnecessary components of their bodily effect's causal histories. The consequence that we do not need to cite mental events to explain the occurrence of their bodily effects is a troublesome result that comes along with our property dualist's position.[117] Therefore, at least part of the

[117] Our property dualist can rightly object that mental events are not entirely explanatorily dispensable with respect to their effects, since we still need to cite them in order to *make sense* of

reason why overdetermination is problematic applies to those nonphysicalists who take mental and neurophysiological events to be linked by contingent, irreducible psychophysical laws. Overall, if a nonphysicalist is to avoid the worries associated with overdetermination, they had either accept epiphenomenalism or deny Completeness.

Section 5.3: Event Realization and Overdetermination

My mental events are *realized* by certain neurophysiological events occurring in my brain and nervous system. I have used the language of realization all throughout this essay and it is high time something be said as to what this relation is. The central concern of this section is to propose an analysis of the realization relation and illustrate how it applies to mental and neurophysiological events. Additionally, I aim to show that this relation can be employed to avoid the problems typically found with standard cases of overdetermination.

Importantly, the realization relation must be spelled out in a way consistent with the kind of *nonreductive physicalism* we have been concerned with in this essay. First of all, realization consists in a nonreductive relation between *events*. In other words, the nonreductive physicalist who accepts that R = realization denies that mental and neurophysiological events are token-identical and they deny this because these events differ in their *de re* modal properties.[118] If mental and

---

our actions. Although the mental is unnecessary to explain why some bodily effect occurs, it still provides a rationalizing explanation for those effects. Perhaps this kind of explanatory power is enough to mitigate the problem with overdetermination that I have argued our property dualist faces. I cannot pursue this response further here, but I believe it is worth consideration.

[118] In Section 1.2.2, we discussed some modal arguments along these general lines: mental event m is essentially F; neurophysiological event n is not essentially F; therefore, m ≠ n (or alternatively:

neurophysiological events are not token-identical, we can ask in what sense they are distinct. As we saw in Section 1.2, distinctness is an ambiguous notion. Events p and q are *weakly modally distinct* in virtue of the possibility that q could occur in the absence of p, although p asymmetrically necessitates q. In Section 5.2.1, we looked at a physicalist mind-body relation (viz., R = determination) of this sort, which holds that it is possible for mental event m to occur in the absence of the neurophysiological event n while n necessitates m. Events p and q are *strongly modally distinct* in virtue of the possibility that either p or q could occur in the absence of the other, which implies that no necessitation relation holds between them. In Section 5.2.2, we looked at a nonphysicalist mind-body relation of this sort, which holds that it is possible for m to occur in the absence of n and for n to occur in the absence of m. In this section, we are concerned with the realization relation which falls somewhere in between these extremes. Where R = realization, mental and neurophysiological events are not token-identical because they are *strongly modally distinct*, but their relationship is not one of brute psychophysical law. Rather, it is explainable in a way acceptable to the physicalist.

Section 5.3.1: The Technical Apparatus

Ever since Putnam's (1975b) proposal of mind-body functionalism, talk of realization has become common philosophical parlance. But only recently has

---

neurophysiological event n is essentially G; mental event m is not essentially G; therefore, n ≠ m). These kinds of modal arguments have been challenged in various ways. See (Lewis 1971), (Gibbard 1975), and (Della Rocca 1996).

there been attempts to analyze this relation.[119] Let us begin by first looking at some cases of event realization. Consider an example inspired by Fodor's (1974) discussion of multiple realizability in which the handing over of particular physical items (e.g., dollar bills) realizes a specific monetary exchange. Presumably, this is so because my handing over these physical items in the right circumstances performs the economic functions of *being a monetary exchange* (e.g., having particular sorts of relations to certain economic institutions, etc.). By performing this function, my handing over of these physical items, in some sense, "makes real" a monetary exchange. Putnam (1975b) also provides an example: some micromechanical event occurring at the hardware-level in my computer's central processing unit is said to realize some computation in virtue of that hardware-level event being in the right circumstances to perform the associated computational functions (e.g., having certain formal/mathematical relations to other "machine states"). When this hardware-level event performs this function, it "makes real" the computation. The fundamental idea of *event realization* seems to be this: one kind of event p realizes another kind of event q when, in certain circumstances, p performs the function associated with q-type events. This intuitive characterization is certainly rough, but it brings out the central components that must be elucidated in our analysis of realization.

The first thing our intuitive characterization makes clear is that realization is a relation that holds between distinct *kinds of events*. Following Lynne Rudder

[119] See (Gillett 2002, 2003), (Polger 2004, 2008), (Shapiro 2000, 2004), and (Shoemaker 2001, 2003c, 2007).

Baker (2002), we can understand these kinds as "primary kinds", where a primary kind is specified by an answer to the question 'What most fundamentally is x?' (Baker 2002, 32). An object's primary kind is essential to that object and among other things determines the persistence conditions of the object. This expresses the idea that objects have certain of their properties essentially and others accidentally. This is easily extended to events, since the question 'What most fundamentally is x?' seems applicable to them just as much as it does to objects. Like objects, events have some of their properties essentially and others accidentally. The monetary exchange that took place between myself and the shopkeeper is most fundamentally *a monetary exchange* and could not have occurred at any time and in any world where it failed to be so. The computation of '2 + 3' is most fundamentally *a computation of '2 + 3'* and, similarly, could not have occurred at any time or in any world where it failed to be such a computation. Let us say, then, that when the question 'What most fundamentally is x?' pertains to events, our answer will be some primary event-kind.

Additionally, our intuitive characterization of realization encourages we make a distinction between different sorts of primary event-kinds. Some primary event-kinds are associated with a function; others are not. Specifically, the realiz*ed* event-kind has a distinctive function that individuates it from other event-kinds. In other words, the primary event-kinds that are realized are *functional event-kinds*, where these primary event-kinds are individuated extrinsically, i.e., in terms of relations to other sorts of things. However, what relations individuate these

functional event-kinds depend on the kind in question. Some functional event-kinds are individuated in terms of *causal dispositions*. This is plausibly the case for the functional event-kind *monetary exchange*, which is individuated in terms of its disposition to cause various sorts of economic events.

Other functional event-kinds are individuated in terms of their *causal history*. For example, Jerrold Levinson (1979) argues that a work of art is essentially historical such that it is "a thing intended for regard-as-a-work-of-art" as works of art have been correctly regarded in the past (Levinson 1979, 234). If this is correct, then the primary event-kind *creation of an artwork* must be a functional event-kind individuated in terms of a specific causal history, namely one involving specific artistic intentions. Another example comes from Ruth Millikan (1989) who argues that hearts are functionally individuated in terms of a particular history of selection. This implies that the primary event-kind *beating of a heart* is a functional event-kind individuated (at least partly) by a specific evolutionary history.

Still other functional event-kinds are individuated in terms of *non-causal relations*. According to Thomas Polger (2008), computations like addition are functional kinds individuated in terms of their "formal or mathematical relations" to other "machine states" (Polger 2008, 240).[120] I want to leave it open that some functional event-kinds could be *non-causal* functional kinds. However, our

[120] One lesson I draw from Polger's discussion is that our analysis of realization should leave open the possibility that not every functional kind is a causal-functional kind. One criticism Polger raises against Gillett's (2002, 2003) account of realization is that it does not keep this possibility open and thus fails to apply to Putnam's (1975b) original example of computations being realized in hardware.

concern with the mental's realization in the neurophysiological will reveal that the functions are causal in one or both of the ways outlined above.

Finally, our intuitive characterization mentions one kind of event performing the function associated with some other event-kind. What this amounts to depends on the functional event-kind in question. If a functional event-kind is individuated in terms of causal dispositions (e.g., monetary exchange), then what it is for something to perform the function associated with *monetary exchange* is for it to possess the causal dispositions which individuate that functional event-kind. When my handing over of particular physical items performs the function associated with *being a monetary exchange*, it is disposed to enter into a specific set of causal relations with other economic events. Now, my handing over of these items need not occur and be so disposed, for it could have occurred in a world where none of the requisite economic institutions exist. The monetary exchange between me and the shopkeeper is *essentially* causally disposed to have certain economic effects, but my handing over of these physical items is not.

If a functional event-kind is individuated in terms of causal history (e.g., creation of an artwork), then what it is for something to perform the function associated with *creation of an artwork* is to have the causal history which individuates that functional event-kind. When the strokes of my paint brush on a blank canvas perform the function associated with *being the creation of an artwork*, its causal history includes a specific artistic intention, namely that the thing being created be regarded as past artworks have been properly regarded. Similarly, these

actions of mine need not occur and have such a history, for they could have occurred in a world in which I had no such artistic intention. My creation of an artwork *essentially* has a particular causal history involving an artistic intention of a certain kind, but the strokes of my paint brush on a blank canvas do not.

Finally, if a functional event-kind is individuated in terms of non-causal relations (e.g., computations), then what it is for something to perform the function associated with *that computation* is to be causally disposed to enter into those relations which "map" onto the formal/mathematical relations which individuate that functional kind. Robert Van Gulick (1988) puts this nicely when he writes that performing the function associated with a computation requires,

> Roughly that there be some mapping from the formal states, inputs, and outputs of the abstract machine table into physical states, inputs, and outputs of the instantiating system, such that under that mapping the relations of temporal sequence among those physical items are isomorphic to the relations for formal succession among the machine table items (Van Gulick 1988, 80).

When some micromechanical event at the hardware level performs the function associated with *being a computation of '2 + 3'*, it is causally disposed to enter into relations that are isomorphic with specific formal/mathematical relations. Again, this event at the hardware level need not be so causally disposed, for it could have occurred outside of my computer's central processing unit on an isolated microchip. The computation of '2 + 3' is *essentially* related to certain other "machine states", but the event at the hardware level is not essentially causally disposed to enter into relations that "map" in the requisite way.

The analysis of realization I offer is inspired partly by Lynne Rudder Baker's (2000, 2002) *modal analysis of constitution* and I want to draw some connections between event realization and constitution. The first reason I find it natural to model our analysis of realization on Baker's account of constitution is that both relations are understood to capture the idea of one kind of thing "making real" another kind of thing. Baker writes,

> Constitution makes an ontological difference. When a piece of marble comes to constitute a statue, it is not just that the piece of marble acquires a new property of being a statue. Rather, a new thing of a new kind with new causal powers and new persistence conditions – a statue – constituted by that piece of marble, comes into existence (Baker 2002, 33).

Likewise, realization makes an *ontological difference*, bringing into being an instance of a new kind of event. These new kinds of events have their own "causal powers" and, at least in some cases, are individuated entirely in terms of those powers. The second reason I find it natural to model realization on Baker's modal analysis of constitution is that the *relata* of both relations are taken to be, in some sense, a "unity" while being distinct, viz., they differ in their *de re* modal properties. According to Baker, constitution is a relation intermediate between identity (in a classical sense) and what she calls "separate existence". She says, "If x and y are constitutionally related at t, there is a unity of x and y at t – a unity without identity" (Baker 2002, 39). The nonreductive physicalism considered here maintains a similar position with regard to, e.g., mental and neurophysiological events. When the mental is realized in the neurophysiological, they *form a unity without*

identity.[121] Now, I want to be clear that Baker's analysis of constitution is extremely controversial and I do not intend anything I say about event realization to commit me to accepting her analysis of the relation between, e.g., the piece of marble and the statue.

These remarks on the similarities between event realization and constitution are programmatic, but I hope they provide at least some motivation for why I have found it fruitful to make connections to Baker's understanding of constitution. There are more concrete similarities I have found between these relations as well. When the piece of marble constitutes the statue, it is commonly held that the lump and statue occupy the same space and share all the same parts. In other words, constitution is standardly taken to require both *spatial* and *material coincidence*.[122] It is common to talk about objects having spatial locations and parts, but it is not as common to think of events in this way.

There are, of course, important differences between objects and events. For instance, as Hacker (1982) has pointed out, objects but not events are properly said to *exist*, while events and not objects are properly said to *occur*. However, this difference seems to be of little importance, since both objects and events are *concrete entities*, existing or occurring in both space and time. Whereas most

[121] The idea of a "unity without identity" is similar to some ideas developed in detail by Yablo (1987). Entities are strictly distinct if they differ in *any* of their properties (e.g., *de re* modal properties), but there is a range of "identity-like" relations which "seem to be ways of being contingently identical" (Yablo 1987, 296). Although I have not worked out the specifics, I think the analysis of event realization offered here can be understood as a way of being contingently identical as Yablo understands this notion.

[122] See (Thomson 1998, 155) and (Doepke 1996, 196 – 198) for a discussion of constitution and spatial coincidence. See (Sider 2002), (Wasserman 2002), and (Zimmerman 1998, 2002) for a discussion of constitution and material coincidence. Mereological notions are notably absent from Baker's (2000, 2002) modal analysis.

objects have fairly clear locations in space, most events do not. The statue is located on the artist's desk and we are easily able to recognize its spatial boundaries through our perceptual faculties. But the artist's creation of the statue, an event, does not have spatial boundaries we can easily ascertain through perception. But it does not follow from this that the event lacks a location in space altogether. In fact, although it is not a straightforward matter to say exactly where the event is located, it does seem relatively clear that it has some location or other. The artist's creation of the statue occurs *in* her art studio and not, for instance, in her home across town. Therefore, even though the spatial location and boundaries of events are unclear to perception and perhaps indeterminate, it stands to reason that *events have locations in space*.

Historically, another important difference between objects and events has been that the former, but not the latter, persist through time by *enduring*. In short, objects have often been thought to be "wholly present" at each moment of their existence, whereas events are said to take up time, have duration, and thus persist by having different temporal parts at different moments of their occurrence.[123] Some philosophers argue that objects, just like events, have temporal parts, but the important point here is that events are standardly taken to be mereologically complex entities. Consider, for example, my 12th birthday party. The party began at noon and ended at 2pm on September 1st, 1996. This means the event that is my 12th birthday party persists through this duration of time entailing that it exists at

[123] See, for example, (Mellor 1980).

1pm of that day. But it is quite absurd to think that my 12th birthday party *endures* through this two hour period. For example, my 12th birthday party is not "wholly present" at 1pm. This is why you could not show up at 1pm and leave a moment later and have stayed for the entire party. Instead, only a part of my 12th birthday party occurs at 1pm, another part occurs at noon, and still another part occurs at 2pm. My 12th birthday party is mereologically complex by being temporally extended and thus has temporal parts. More generally, we can say that all events are mereologically complex entities having temporal parts.

Now, consider the monetary exchange between me and the shopkeeper which occurred during some interval of time. This event is said to be realized by my handing over of particular physical items. It strikes me as *prima face* plausible that this economic event occurred in the same location as my handing over of these items. Furthermore, it is also plausible that these events occurred during the same duration of time. How long did the monetary exchange take? As long as it took me to hand over the physical items. The same, I believe, can be said of our other cases of event realization. The computation of '2 + 3' is realized by some micromechanical event in my computer's central processing unit and, plausibly, these events occurred in the same location and during the same interval of time. The computation took place in my computer's central processing unit right where its realizer occurred and its temporal boundaries are the same as its realizer's temporal boundaries. Moreover, this would seem to be precisely what the nonreductive physicalist would want to say about my belief that water is wet and

the pattern of neural firings in my prefrontal cortex. My belief has a location in space and it is located precisely where its realizer is. Additionally, my belief (at least as long as it is an *occurrent belief*) would seem to have the same duration as its neurophysiological realizer. Therefore, it is a plausible condition on an analysis of event realization that the realized and realizing events are *spatially* and *temporally coincident*.

The final similarity I want to point out between event realization and constitution is an appeal to circumstances. Above I claimed that an intuitive characterization of event realization is that one kind of event realizes another kind of event when, *in particular circumstances*, the first event performs the function associated with the second event. What are these circumstances supposed to be? Following Baker, we can say that these circumstances are the "milieu" in which something can perform the function which individuates the functional event-kind. However, we must put an obvious restriction on our characterization of these circumstances. They cannot themselves entail the occurrence of the functional event-kind in question. These circumstances must be necessary, but not sufficient, conditions for the functional event-kind to occur then and there (Baker 2000, 42).

Moreover, the correct characterization of this "milieu" depends on the correct theory of that functional event-kind. For instance, the circumstances in which some micromechanical event in my computer's central processing unit realizes an addition computation will be different from the circumstances in which my handing over particular physical items realizes a monetary exchange, for the

function associated with monetary exchanges is different than the function associated with computations. What are the circumstances in which my handing over of particular physical items can realize a monetary exchange? These circumstances entail the existence of intelligent creatures with certain kinds of intentions, the existence of particular social and economic institutions, and perhaps even certain conventions. What are the circumstances in which a micromechanical event in my computer's central processing unit realizes a computation of '2 + 3'? These circumstances entail the existence of a particular sort of computational architecture being implemented in the computer hardware, an architecture that sustains the potential to enter into the requisite mathematical and formal "machine states". Roughly, the role of circumstances here is to "enable" the realiz*ing* event to perform the function associated with realiz*ed* primary event-kind.

Section 5.3.2: A Modal Analysis of Event Realization

Now that some of the technical apparatus is in hand, I propose the following analysis of event realization inspired by Baker's (2000, 2002) account of constitution:

(ER) An event p realizes an event q during t iff there are distinct primary event-kinds P and Q and Q-favorable circumstances such that:

(1) p has P as its primary event-kind,
(2) q has Q as its primary event-kind, where Q is a functional kind individuated in terms of some function F,
(3) p and q are spatially and temporally coincident during t,
(4) p occurs in Q-favorable circumstances during t,
(5) necessarily: for any event e and duration t, if e has P as its primary event-kind and e occurs in Q-favorable circumstances during t, then there occurs some event g such that g has the functional event-kind Q as its primary event-kind and g is spatially and temporally coincident with e,

(6) possibly: there is a duration t such that p occurs and there is no h such that h has the functional event-kind Q as its primary event-kind and h is spatially and temporally coincident with p.

I have already discussed conditions (1) – (3) above. Condition (4) introduces the notion, following Baker, of "Q-favorable circumstances". Above I glossed this notion as the "milieu" which is necessary, but not sufficient, for the occurrence of the functional event-kind Q. Recall that the proper characterization of the Q-favorable circumstances depends on the correct theory of the functional event-kind. Condition (5) states that it is a necessary truth that when a P-event occurs in Q-favorable circumstances, there occurs a Q-event which is spatially and temporally coincident with the P-event. Condition (6) entails that the P-event is not identical to the Q-event, since it is possible for the P-event to occur in the absence of any Q-type event spatially and temporally coincident with it.

The nonreductive physicalism we are primarily concerned with in this section claims that mental events are realized in neurophysiological events. According to (ER) this implies that: the pattern of neural firings in my prefrontal cortex has as its primary event-kind some neurophysiological kind N; my belief that water is wet has as its primary event-kind a mental kind M, which is distinct from N and individuated in terms of some function; the neural firings in my prefrontal cortex and my belief that water is wet are spatially and temporally coincident; these neural firings occur in M-favorable circumstances; it is a necessary truth that when an N-event (like these neural firings) occur in M-favorable circumstances, there occurs an M-event (like my belief that water is wet)

spatially and temporally coincident with it; and, finally, it is possible that the N-event occurs in the absence of any M-type event spatially and temporally coincident with it. There are two features of this application of (ER) that I want to discuss: that M is a primary event-kind individuated in terms of some function and that these neural firings are in M-favorable circumstances. These features are connected and I want to illustrate just how they are *vis-à-vis* our present case.

Let us begin by considering an objection raised by Derk Pereboom (2002). His discussion suggests that our condition (6) is too strong. Pereboom's principal concern is that there could be cases in which a neurophysiological event n realizes a mental event m where "[n] could exist without being spatially coincident with [m] but not without [n] being spatially coincident with something of the same primary kind as [m]" (Pereboom 2002, 616). Suppose that the primary event-kind of my belief that water is wet is *belief*. He continues,

> It would seem that a token neural state could exist without being spatially coincident with the belief that water is wet – say, on Twin Earth – nevertheless it must be that this neural state be spatially coincident with some belief or other – on Twin Earth it would be spatially coincident with the belief that twin-water is wet (Pereboom 2002, 616).

It seems to me that Pereboom is wrong that the "neural state" which realizes my belief that water is wet must be spatially coincident with some belief or other. Is it not a genuine possibility that this neurophysiological event could occur outside of its actual neural context in some isolated brain matter afloat in agar jelly? In other words, it seems to me that we can imagine this very pattern of neural firings in my prefrontal cortex as occurring in circumstances far removed its actual

circumstances. In such a radically different environment, no coincident belief of any kind occurs.

However, Pereboom's concern raises an important point. We must recognize that the primary event-kind of my belief that water is wet is not simply *belief*. When we ask 'What most fundamentally is x?' about some intentional mental event such as my belief that water is wet, it seems correct to say that it most fundamentally is a *belief*. There is no time or world in which that very event occurs and fails to be a belief. This event could not, for example, have been a desire or a pain. But it also seems correct to say that it is most fundamentally a contentful event related to a particular propositional content, namely *that water is wet*. At no time and in no world could this belief have occurred and failed to have this propositional content. A belief that twin-water is wet would not have been my belief. Therefore, when the question 'What most fundamentally is x?' pertains to intentional mental events, like my belief that water is wet, our answer must specify an *intentional event-kind*, which is an ordered pair of an *attitude event-kind* (e.g., belief, desire, intention, etc.) and a *propositional content*. On Twin-Earth, the neurophysiological event n may be coincident with some belief or other, but it is wrong to say that it is coincident with some event of the relevant *intentional event-kind*, which specifies both an attitude event-kind and a propositional content. When we apply (ER) to intentional mental events, we must keep in mind that their primary event-kinds are intentional event-kinds and not merely attitude event-kinds.

If intentional mental events are realized in neurophysiological events, then, by condition (2), intentional event-kinds must be associated with a particular function. What are these functions? And what are the M-favorable circumstances in which some neurophysiological event is said to "perform" that function? Without answers to these questions our account of the mental's realization in the neurophysiological will be incomplete. One virtue of (ER) is that it connects, in a rather intimate way, with much of the theorizing done on *mental content*. In particular, (ER) enables us to see that different theories of mental content are attempting to answer our two questions above. Another virtue of (ER) is that it is consistent with either individualistic or anti-individualistic theories.

According to individualistic theories, the functions associated with intentional event-kinds are relations entirely *internal* to the psychological subject. One influential version of this individualism is proposed in Block (1986). What is the function that individuates the intentional event-kind belief that water is wet from other intentional event-kinds? The answer appeals to the inferential or conceptual role of this type of event in the subject's psychology. More specifically, Block writes that the function which individuates intentional event-kinds is

> The causal role of the [event-kind] in *reasoning and deliberation* and, in general, in the way the [event-kind] combines and interacts with other [event-kinds] so as to mediate between sensory inputs and behavioral outputs (my emphasis) (Block 1986, 93).

In short, this theory specifies the function associated with intentional event-kinds in terms of a specific set of *causal dispositions*. What makes some mental event a belief that water is wet are the causal dispositions it possesses to other sorts of

intentional mental events in the subject's reasoning and deliberation. Now, if this individualistic theory is correct, the M-favorable circumstances for the intentional event-kind belief that water is wet can be characterized by a list of open sentences describing the causal dispositions that individuate that event-kind:[124]

> x is disposed to cause beliefs that water is good for washing clothes
> x is disposed to cause beliefs that water is liquid at room temperature
> x is disposed to cause water seeking behavior given certain sorts of desires
> x is disposed to be caused by certain sorts of perceptual experiences
> etc.

When this list of open sentences is true of some neurophysiological event n, it is in the relevant M-favorable circumstances and is said to "perform" the function associated with the intentional event-kind. By condition (5) in (ER), if n is in M-favorable circumstances, there occurs a belief that water is wet which is spatially and temporally coincident with n.

A venerable tradition stemming from arguments given by Putnam (1975a) and Burge (1979, 1986) has it that the functions associated with intentional event-kinds are *not* entirely relations internal to the psychological subject. These anti-individualistic theories claim that the functions which individuate intentional event-kinds include, in addition to some set of causal dispositions, relations *external* to the psychological subject. Putnam's (1975a) arguments suggest that in addition to a particular "causal-functional role", intentional event-kinds are individuated in terms of their *causal history* to a particular physical environment. Specifically, the psychological subject must be "embedded" in an environment *in which there is*

[124] See (Baker 2002, 44) for characterizing circumstances in terms of open sentences.

water in order to have, e.g., a belief that water is wet. Burge (1979, 1986) extended this line of reasoning by including a causal history within a particular socio-linguistic community. These anti-individualistic theories specify the functions associated with intentional event-kinds in terms of a set of causal dispositions *plus* causal-historical relations to a physical and/or social-linguistic environment. If these sorts of anti-individualistic theories are correct, the M-favorable circumstances for the intentional event-kind belief that water is wet can be characterized by a list of open sentences describing a set of causal dispositions and the requisite historical relations that individuate the event-kind:

> x is disposed to cause such and such beliefs
> x is disposed to cause such and such actions given certain desires
> x is disposed to be caused by certain sorts of perceptual experiences
> x is historically related in the right way to a its physical and/or socio-linguistic environment
> etc.

Similarly, when this list of open sentences is true of some neurophysiological event n, it is in the relevant M-favorable circumstances and, by condition (5), is spatially and temporally coincident with a belief that water is wet.

Of course, there is no philosophical consensus on precisely what the correct theory of mental content is, but central to our present project is that any of these theories individuate different kinds of mental events in relational terms.[125] Either some set of causal dispositions or a set of dispositions *plus* a causal history individuates mental event-kinds. The best theories we have on mental content

[125] Phenomenal events, viz., "qualia", are perhaps a different story and arguably are not individuated in terms of a function in the sense at issue. If this is correct, then there is a subset of mental event-kinds which are not able to be realized in the neurophysiological. I have my doubts, but will not pursue this vexed topic here.

entail that these kinds of events are prime candidates for being individuated in terms of functions. Furthermore, each comes along with a story of what it would be for some neurophysiological event to "perform" the function which individuates a particular mental event-kind. In other words, a theory of mental content proposes answers to our two questions: (a) what are the functions that individuate intentional event-kinds?; and (b) what are the M-favorable circumstances of some intentional event-kind in which something can realize that event-kind?

Section 5.3.3: Handling Cases of MN-Overdetermination

Proposing an analysis of event realization has been a means to an end. What we really want is a plausible account of a mind-body relation consistent with the nonreductive physicalism under consideration that avoids the problem of causal overdetermination. In the rest of this section, I aim to demonstrate that when R = event realization (along with some further assumptions), MN-overdetermination does not inherit any of the problems associated standard cases of overdetermination. Let's begin with the simplest problem to handle, viz., coincidence and conspiracy.

In Section 5.1.2, we saw that one issue with standard cases of overdetermination is that the causes are correlated with one another in a way that has no explanation or in virtue of some conspiracy. If MN-overdetermination inherits these features, then its systematic nature entails that coincidence or conspiracy are pervasive features of our world. The nonreductive physicalist who endorses that R = event realization can provide a similar response to this concern as Yablo and the property dualist. When the mind-body relation is determination, the

correlation between mental and neurophysiological events is governed by a *metaphysically necessary truth*. Necessarily, when a determinate event occurs, its determinable event also occurs. When the mind-body relation is a brute psychophysical law, the correlation between these events is governed by a *nomologically necessary truth*. Across all possible worlds with our laws of nature, if there occurs a neurophysiological event of a certain sort, a mental event of a specific kind occurs as well. As Sider (2003) points out, "It is no coincidence that … mental and physical events are correlated, given the necessary truths governing these correlations" (Sider 2003, 722).

Similarly, when the mind-body relation is event realization, mental events are correlated with neurophysiological events and this correlation is systematic, but it is *not* a correlation without an explanation nor is it some sort of Leibnizian pre-established harmony. Why? Because there is a *metaphysically necessary truth* that governs this correlation. Necessarily, when a neurophysiological event occurs *in the right circumstances*, viz., properly causally "wired" within the neural environment of an organism embedded in a particular physical and/or socio-linguistic environment, then there occurs a coincident mental event. In short, mental causes are correlated with neurophysiological causes in virtue of the mental's *realization* in our neurophysiology. It is no coincidence or conspiracy that mind and brain make a *concurrent* difference to our bodily movements.

In Section 5.1.3, we discussed the concern that MN-overdetermination entails that overdetermined bodily effects receive a "double dose" of "causal

oomph!". Additionally, this implies that there is a systematic failure of a particular kind of physical explanation. In order to explain why some bodily movement has as much of some physical quantity as it does, we must make an essential appeal to an irreducible mental event. I propose that the best way to handle this particular worry with MN-overdetermination is not to present a solution to the problem, but to "dissolve" it and show that it is not really a problem at all.

In Chapter 4, I argued for the view I called counterfactualism: counterfactual dependence between distinct events is sufficient for causation. Furthermore, I concluded that the main rival to counterfactualism – productive views of causation – are at a significant disadvantage in their attempts to ground mental causation. In short, I provided reason to believe that their requirements on causation, viz., some sort of *physical connection* between cause and effect, is inconsistent with both mental and neurophysiological causation. The empirical details about how mental and neurophysiological events are "hooked up" to the physiology of the human organism precludes a physical connection between cause and effect. Both mental and neurophysiological causes make a difference to our bodily movements in virtue of *disconnecting* a physiological process in our bodies; not in virtue of a *connecting* process.

If these arguments are successful, we have principled grounds on which to "dissolve" this second worry about MN-overdetermination. The problem presupposes that causation involves the transfer of some "causal oomph!" (e.g., energy-momentum) from cause to effect and this sort of transfer requires a physical

connection between the causal relata. The arguments of Chapter 4 concluded that neither mental nor neurophysiological causes are physically connected to their bodily effects. *Ergo*, neither mental nor neurophysiological causation involve the transfer of some "causal oomph!" from cause to effect. My conclusion, then, is that this concern about MN-overdetermination presupposes a false conception of mental and neurophysiological causation. Once we dispel this misguided picture, we can see that "the problem" is not really a problem after all.

The final and most difficult concern is the problem that MN-overdetermination makes both mental events and their neurophysiological realizers causally dispensable. What it means for an event to be causally dispensable *vis-à-vis* some effect is for the following counterfactual to be true of that event: if the event c had not occurred, then the effect e would still have occurred. In other words, overdetermined effects are not counterfactually dependent on their overdetermining causes. This is plausibly true of most causes in standard cases of overdetermination. If MN-overdetermination inherits this problem, then both mental events and their neurophysiological realizers end up being *redundant* and *superfluous* parts of their effect's causal history. This would be a *prima facie* troubling result for the nonreductive physicalist. However, when R = event realization, neither the mental nor its neurophysiological realizer are causally dispensable with respect to their overdetermined bodily effect.

In order to demonstrate this, I shall show that these effects do indeed counterfactually depend on their mental and neurophysiological causes. Let us first consider whether the following counterfactual is true:

(A) If n had not occurred, I would not have searched for something to drink

where n is some neurophysiological event occurring in my brain which realizes my desire m that I drink some water. Our question is this: if n realizes m, then is the *most similar* not-n world a world where I do not search for something to drink? If the most similar world in which n fails to occur is a world where n is *replaced* by a similar neurophysiological event n* which also realizes m, then the counterfactual (A) is false and n is causally dispensable with respect to my searching behavior. But the most similar not-n world is *not* a world where n is "replaced" by n*. These sorts of "replacement readings" are inappropriate in causal contexts. When we imagine n gone, we imagine it gone *punkt*.

Now, properly evaluating (A) must accord with Lewis's similarity metric, which instructs us to tolerate an inconspicuous violation of law in order to match matters of particular fact as much as possible. In other words, we hold the past fixed as much as we can and posit a "small, local miracle" that results in the non-occurrence of the neurophysiological event n. This implies that the most similar not-n world is one where the miracle violates the causal law between n and one of its proximate causal antecedents. This inconspicuous violation of law leaves the M-favorable circumstances in which n actually occurs virtually undisturbed. But, recall, these circumstances are necessary, but not sufficient, for the occurrence of m.

Therefore, in the absence of n, the mental event m also fails to occur. And if the most similar not-n world is a not-m world, then it is also a world where I do not search for something to drink. *Ergo*: the counterfactual (A) is true and n is *not* a causally dispensable part of its bodily effect's causal history.

Now consider whether the following counterfactual is true:

(B) If m had not occurred, I would not have searched for something to drink.

Our question is this: if m is realized by n, then is the *most similar* not-m world a world where I do not search for something to drink? What if the most similar world in which m fails to occur is a world where m is replaced by a similar mental event m* (i.e., a desire that I drink some *twin-water*)? The neurophysiological event n still occurs in this world, but in slightly different circumstances, viz., M*-favorable circumstances instead of M-favorable circumstances. This not-m world is a world where n still occurs and so my searching behavior still occurs. Hence, the counterfactual (B) is false and m is causally dispensable with respect to my searching for something to drink. But it should be obvious by now that a "replacement reading" is not the appropriate way to evaluate (B) in causal contexts. When we imagine that the mental event m fails to occur, we do not replace it with any similar event. As Bennett's (2003) remarks vividly illustrate, "You simply snip it away as though you had a metaphysical hole-puncher" (Bennett 2003, 482).

How, then, do we wield Bennett's "metaphysical hole puncher"? What precisely are the most similar worlds supposed to look in which the mental event m fails to occur? Given that m is *realized* by n, there are two relevant sets of

possibilities to consider. The first consists in those worlds where condition (4) is violated such that the neurophysiological event n does not occur *in the M-favorable circumstances* in which it actually occurs. Depending on the details of the function associated with the intentional event-kind M, this involves changing facts about n's *causal history* and/or n's *causal dispositions*. On the one hand, if n has a causal history different from its actual one, then n occurs in my brain such that I have been causally-historically "embedded" in a different physical and/or socio-linguistic environment. This means these counterfactual worlds diverge from the actual world in numerous *matters of particular fact*. For example, all the actual facts about my past causal interactions with water fail to hold as these interactions have instead been with a different substance, like XYZ. On the other hand, if n has a sufficiently different set of causal dispositions such that it is not capable of performing the causal role which individuates M, then n occurs but is "embedded" in a different neural environment, where it is not "wired" to the rest of my brain in the way it is in the actual world. Again, this implies that our counterfactual world diverges in numerous *matters of particular fact*, viz., facts about my brain's internal neurophysiological connections.

The second set of worlds to consider are those in which condition (4) is violated, not because n fails to occur *in M-favorable circumstances*, but rather because the neurophysiological event fails to occur *simpliciter*. This non-occurrence of n is brought about by an inconspicuous violation of the causal law connecting n with one of its proximate causal antecedents. This leaves the M-

favorable circumstances virtually untouched, but, again, this "milieu" is only necessary and not sufficient for the occurrence of m. Importantly, if m's non-occurrence is achieved in this way, there is an extensive match of particular matters of fact between these worlds and the actual world.

The point of this comparison should be obvious. According to Lewis's similarity metric, the most similar worlds consist of those with a greater match of particular fact achieved at the expense of a small, inconspicuous violation of law. The first set of worlds involves no violation of law, but achieves m's non-occurrence by an extensive mismatch of particular matters of fact. The second set of worlds involves an inconspicuous violation of law and retains an almost perfect match of particular matters of fact. Therefore, the set of most similar worlds consists of this second set of worlds, where m fails to occur because its realizer n fails to occur. If neither m nor n occur in this set of worlds, then my searching behavior also fails to occur. We can conclude, then, that the (B) counterfactual is true and m is *not* a causally dispensable part of its overdetermined effect's causal history.

What this demonstrates is that if the mental event m is realized by the neurophysiological event n along the lines of (ER), then neither event is a causally dispensable part of its overdetermined bodily effect's causal history. The effect is *counterfactually dependent* on both overdetermining causes. Although standard examples of overdetermination result in both events being superfluous or redundant causes, this does not hold true for cases of MN-overdetermination, where R = event

realization. The nonreductive physicalist who upholds that mental events are realized in neurophysiological events can successfully avoid the threat posed by causal dispensability.

I want to end this section with some programmatic remarks on the idea that realization implies a "unity without identity". What I have tried to show is that the (A) and (B) counterfactuals are both true, that is, neither the mental event m nor its neurophysiological realizer n are causally dispensable *vis-à-vis* my searching behavior. However, it is not just that both of these counterfactuals are true, but the set of most similar worlds for each are *exactly the same set of worlds*. Which worlds are relevant to the evaluation of (A) 'if n had not occurred, I would not have searched for something to drink'? The worlds in which an inconspicuous violation of law "breaks" the connection between the event n and one of its proximate causal antecedents. Which worlds are relevant to the evaluation of (B) 'if m had not occurred, I would not have searched for something to drink'? The exact same set of worlds: ones where an inconspicuous violation of law "breaks" the connection between the event n and one of its proximate causal antecedents.

This fact speaks to the almost paradoxical relationship realization ties between realiz*ed* events and their realiz*ing* events.[126] Mental events and their neurophysiological realizers are distinct events. In fact, these events are *strongly modally distinct*, since neither metaphysically necessitates the other. In this way, realization is similar to the relation the property dualist holds between mental and

[126] See (Baker 2002, 38 – 40) for similar remarks concerning constitution and (Yablo 1987) for these sorts of remarks more generally.

neurophysiological events. For instance, if n realizes m, then these events differ in their *de re* modal properties. Yet, these events are so closely related that the very set of worlds relevant to evaluating whether an effect counterfactually depends on m is the same set of worlds relevant to evaluating whether this effect counterfactually depends on n. In this way, realization is similar to the relation the token-reductionist holds between mental and neurophysiological events (viz., token-identity). If n realizes m, then whatever counterfactually depends on m also counterfactually depends on n.

In Chapter 4, I argued for the position that counterfactual dependence between distinct events is sufficient for causation. This conceptual connection implies that if whatever counterfactually depends on m also counterfactually depends on n, then whatever m causes, n causes and *vice versa*. To put it somewhat suggestively, if we had to consider these events *qua causes* only across the set of worlds relevant to establishing counterfactual dependencies, then mental events and their neurophysiological realizers could be treated *as if* they were the very same event.[127] Although it is a relation between *distinct* events, realization binds these events together to form a "causal unity". To my mind, realization presents the physicalist with a viable middle position between the identity of the token-reductionist and the dualism of the property dualist; a middle position that

[127] In Section 5.2.1, we made the same observation with respect to determinates and their determinables. There I said that the Lewisian analog of an intervention – the "small, local miracle" – could not "surgically intervene" on the two events across the set of most similar worlds. This applies to our present case as well.

embraces an MN-overdetermination that inherits *none* of the problems accompanying standard cases of overdetermination.

Conclusion

In his discussion of Descartes's substance dualism, Gassendi wondered, "How could there be effort directed against anything, or motion set up in it, unless there is mutual contact between what moves and what is moved? And how can there be contact without a body … ?".[128] Gassendi's conception of causation as involving contact between cause and effect is certainly antiquated, but his question is penetrating and resonates to this day. How could the mental cause anything *unless* it just is something physical? If the mental isn't physical, then it must either be epiphenomenal or a merely redundant, overdetermining cause. This is the dilemma we have been concerned with throughout this essay. I have attempted a systematic treatment of this problem from the perspective of the nonreductive physicalist, those who deny that the mental "just is" something physical.

In Chapter 2, I discussed several important responses to resolving this dilemma in order to differentiate my preferred solution from them. One could deny the reality of mental phenomena in the way advanced by John Heil (1999, 2003). Or one could reject the Homogeneity Assumption identified by Tim Crane (1995) and claim that mental causation is a different sort of causation from neurophysiological causation. Lynne Rudder Baker (1993) rejects the assumption of Completeness, thereby dismantling the dilemma before it even gets started. Fred

[128] *The Essential Descartes*, ed. M. Wilson (New York: New American Library, 1969): 373.

Dretske (1988) and Ausonio Marras (1998) argue, in different ways, that mental phenomena have different effects than neurophysiological phenomena. Finally, one could simply embrace Jaegwon Kim's preferred solution and reduce the mental to the neurophysiological in order to avoid the dilemma's horns. My rejection of these responses define the "problem space" in which I provide my preferred solution to the exclusion problem. In other words, I assume throughout this essay that mental phenomena are real, that mental causation is the same sort of causation as neurophysiological causation, that Completeness is true, that mental phenomena have the same effects as neurophysiological phenomena, and that both type- and token-reductionism are false.

In Chapter 3, I discussed and criticized Stephen Yablo's and Sydney Shoemaker's attempts to solve the exclusion dilemma. Their common approach appeals to an intimate relation between mental and neurophysiological events, and construes making a causal difference as satisfying a proportionality constraint. Their approach to the problem has much to recommend it, but I argue that the proportionality requirement on causation cannot withstand critique. If the proportionality constraint is consistently applied, it leaves few pre-reflective causal judgments intact and, therefore, ought to be rejected. This motivates looking for a different account of what it is to make a causal difference, one that better matches our intuitive judgments about causation.

In Chapter 4, I argued that we should understand what it is to make a causal difference in terms of counterfactual dependence between distinct events and that

this account can vindicate mental causation. Moreover, I defended this position, which I called *counterfactualism*, from some recent criticisms made by Jaegwon Kim. According to Kim, counterfactual dependence cannot vindicate mental causation because it cannot adequately distinguish between genuine causal relations and pseudo-causal relations nor does it satisfactorily ground human agency. Kim's underlying motivation for rejecting counterfactualism stems from the *production intuition*, the idea that causation involves spatiotemporal local and contiguous processes connecting causes with their effects. This alternative to counterfactualism fails, however, for it is inconsistent with the physiological mechanisms of human action. Counterfactualism, or something very similar to it, remains our only option for vindicating the efficacy of our beliefs and desires.

The positions defended in previous chapters forced me to address the horn of overdetermination, which I undertook in Chapter 5. I embraced the consequence that our bodily movements are overdetermined by both a mental and neurophysiological cause, but argued that this sort of overdetermination, which I called MN-overdetermination, is entirely unproblematic. To my mind, there are three reasons why overdetermination is a troublesome consequence and none of these reasons apply to cases of MN-overdetermination. That is, none of them apply if (a) counterfactualism is endorsed and (b) the relation between mental and neurophysiological events is understood along lines of (ER).

However, one final question remains: which of the assumptions that generate the exclusion problem do I reject? I believe the problem lies in the

implicit assumption that our bodily movements are *not* causally overdetermined. Proponents of the exclusion problem, like Kim, go awry when they assume that all cases of overdetermination are problematic and ought to be avoided. But what I have argued in Chapter 5 is that there are some cases of overdetermination – cases that systematically occur in our world no less! – that are no cause for concern. When the mental is realized in the neurophysiological in accordance with (ER) and mental causation is grounded in counterfactual dependence, our reasons for finding overdetermination *bad* simply do not apply.

One upshot of our discussion is that the so-called "causal argument" for reductive brands of physicalism either begs the question against the nonreductionist or includes an unjustified premise. In Section 1.5.1, we saw the following kind of argument, which made an essential appeal to an "overdetermination is bad" premise:

(P1) For every physical event p that has a sufficient cause occurring at t, some physical event p* is causally sufficient for p at t,

(P2) All mental phenomena have physical effects,

(P3) The physical effects of mental events are not overdetermined,

(P4) Therefore, mental events must be identical with physical events.

The third premise clearly involves the implicit assumption that overdetermination is something to be avoided. Perhaps because overdetermination is "pre-theoretically … an ugly picture" (Merricks 2001, 67) and it is "at best extremely odd to think that each and every bit of action we perform is overdetermined in

virtue of having two distinct sufficient causes" (Kim 1989, 86). But without providing some reason why overdetermination is bad, (P3) simply begs the question against the *nonreductive* brand of physicalism endorsed here.

Perhaps overdetermination is bad for some unstated reason. If so, then proponents of the "causal argument" can justify (P3) by appealing to these sorts of reasons. In the last chapter, I considered three different ways of justifying (P3), none of which apply to the overdetermination embraced by the counterfactualist who maintains that mental events are *realized* in neurophysiological events. Therefore, until some further justification is produced, we have no reason to accept (P3) and *a fortiori* no reason to accept the reductive conclusion of the "causal argument".

Does this mean the "causal argument" is entirely bankrupt? I do not think this follows. For a similar sort of argument still provides some grounds on which to reject the nonreductive physicalisms that insist on *productive mental causation* or property dualisms which hold that contingent, but fundamental, psychophysical laws connect mental and neurophysiological events. If causation involves a physical connection between the causal relata, where some "oomph!" is transferred in the process, then MN-overdetermination entails a duplicative transfer of "causal oomph!" from which it follows that a certain sort of physical explanation systematically fails in our world. Additionally, if a fundamental psychophysical law connects mental and neurophysiological events, then MN-overdetermination entails that mental events are causally dispensable parts of their effect's causal

history. These are both reasons to accept (P3) and move back towards token-reductionism or, alternatively, abandon productive mental causation or property dualism, respectively.[129]

Mental causation holds a primary place in our conception of ourselves as persons. We are profoundly special. We can be properly praised and blamed for our actions and our choices. We are the proper targets of a dazzling array of moral attitudes. We direct these attitudes at others and at our own selves. And all of this makes sense only because our reasons and motives, our beliefs and our desires, our self-reflections and deliberations have the power to bring about what we do and the choices we make. But we are also human beings, subject to the same physical laws and amenable to the same physical explanations as every other complex system in world. We are profoundly special, but not exceptions.

But our personhood and our creature-hood struggle to sleep comfortably with one another. The exclusion problem is just one of the many problems that arise from the apparent conflict between the manifest image of ourselves as *persons* and the scientific image of ourselves as *biological organisms*. Most philosophers are, by nature, reconciliatory. This essay represents my attempt to reconcile the mental causation which underlies our personhood with the neurophysiological causation that underlies our creature-hood.

---

[129] In any case, I think there are reasons to abandon productive mental causation and property dualism independent of the above sort of "causal arguments". We have seen the reasons to reject productive mental causation in Section 4.2.3. See (Bennett 2005) for some reasons to doubt property dualism.

Bibliography

Anscombe, G. E. M. (1975). "Causality and Determination", *Causality and Conditionals*, Ernest Sosa (ed.) (Oxford: Oxford University Press): 63 – 81.

Armstrong, D. (1981/2002). "The Nature of Mind", *Readings in the Philosophy of Psychology*, vol. 1 (Cambridge: Harvard University Press). Reprinted in *Philosophy of Mind: Classical and Contemporary Readings*, D. Chalmers (ed.): 80 – 87.

Aronson, J. (1971). "On the Grammar of 'Cause'", *Synthese* 22: 414 – 430.

Baker, L.R. (1993). "Metaphysics and Mental Causation", *Mental Causation*, John Heil and Alfred Mele (eds.) (Oxford University Press): 75 – 95.

Baker, L.R. (1997). "Why Constitution is not Identity", *Journal of Philosophy* 94: 599 – 621.

Baker, L.R. (2000). *Persons and Bodies: A Constitution View* (Cambridge: Cambridge University Press).

Baker, L.R. (2002). "Making Things Up", *Philosophical Topics* 30: 31 – 51.

Bennett, K. (2003). "Why the Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It", *Noûs* 37: 471-497.

Bennett, K. (2005). "Why I am not a Dualist". Retrieved March 23rd, 2013 from David Chalmers's website: http://consc.net/online/1/all#.1.4f

Bennett, K. (2008). "Exclusion Again", *Being Reduced: New Essays on Reduction, Explanation, and Causation*, Hohwy and Kallestrup (eds.) (Oxford: Oxford University Press): 280 – 305.

Bennett, J. (1988). *Events and Their Names*. (Indianapolis: Hackett).

Bennett, J. (2002). "What Events Are", *The Blackwell Guide to Metaphysics*, Gale (ed.) (Blackwell): 43 -65.

Bickle, J. (1997). *Psychoneural Reductionism: The New Wave* (MIT Press).

Block, N. (1980). "Troubles with Functionalism," *Readings in the Philosophy of Psychology*, vol. 1 (Cambridge: Harvard University Press).

Block, N. (1986). “Advertisement for a Semantics for Psychology”, *Midwest Studies in Philosophy* 10 : 615 – 678.

Block, N. (1990). “Can the Mind Change the World?”, *Meaning and Method*, George Boolos (ed.) (Cambridge: Cambridge University Press).

Block, N. (2003). “Do Causal Powers Drain Away?”, *Philosophy and Phenomenological Research* 67: 133 – 150.

Block, N. and Fodor, J. (1972). “What Psychological States are Not,” *Philosophical Review* 81: 159-181.

Block, N. and Stalnaker, R. (1999). “Conceptual Analysis, Dualism, and the Explanatory Gap”, *The Philosophical Review* 108: 1 – 46.

Bontly, T. (2005). “Proportionality, Causation, and Exclusion”, *Philosophia* 32: 331 – 348.

Burge (1979). “Individualism and the Mental”, *Midwest Studies in Philosophy* 4: 73 – 121.

Burge (1986). “Intellectual Norms and the Foundations of Mind”, *The Journal of Philosophy* 83: 697 – 720.

Castaneda, H.N. (1984). “Causes, Causity, and Energy”, *Midwest Studies in Philosophy*, P. French, T. Uehling, and H. Wettstein (eds.) (Minneapolis: University of Minnesota Press): 17 – 27.

Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. (Oxford: Oxford University Press).

Chalmers, D. (2002). “Consciousness and Its Place in Nature,” *Blackwell Guide to the Philosophy of Mind*, S. Stich and T. Warfield (eds.) (Blackwell).

Crane, T. (1992). “Mental Causation and Mental Reality”, *Proceedings of the Aristotelian Society* 92: 185 – 202.

Crane, T. (1995). “The Mental Causation Debate”, *Proceedings of the Aristotelian Society, Supplementary Volumes* 69: 211 – 236.

Crane, T. (2001). *The Elements of Mind* (Oxford: Oxford University Press).

Crisp, T. and Warfield, T. (2001). “Kim’s Master Argument”, *Noûs* 35: 304-316.

Davidson, D. (1980a). “Actions, Reasons, and Causes”, *Essays on Actions and Events* (Oxford: Oxford University Press): 3 – 19.

Davidson, D. (1980b). “Causal Relations”, *Essays on Actions and Events* (Oxford: Oxford University Press): 149 – 162.

Davidson, D. (1980c). “The Individuation of Events”, *Essays on Actions and Events* (Oxford: Oxford University Press): 163 – 180.

Davidson, D. (1980d). “Mental Events,” *Essays on Actions and Events* (Oxford: Oxford University Press): 207 – 227.

Davidson, D. (1980e). “Psychology as Philosophy,” *Essays on Actions and Events* (Oxford: Oxford University Press): 229 – 238.

Davidson, D. (1980f). “The Material Mind”, *Essays on Actions and Events* (Oxford: Oxford University Press): 245 – 260.

Davidson, D. (1993). “Thinking Causes”, *Mental Causation*, John Heil and Alfred Mele (eds.) (Oxford University Press): 3 – 17.

Della Rocca, M. (1996). “Essentialists and Essentialism”, *Journal of Philosophy* 4: 186 – 202.

Doepke, F. (1997). “Spatially Coinciding Objects”, *Material Constitution: A Reader*, Michael Rea (ed.) (Oxford: Rowman & Littefield Publishers): 10 – 24.

Dowe, P. (1992). “Wesley Salmon’s Process Theory of Causality and the Conserved Quantity Theory”, *Philosophy of Science* 59: 195 – 216.

Dowe, P. (1995). “Causality and Conserved Quantities: A Reply to Salmon”, *Philosophy of Science* 62: 321 – 333.

Dowe, P. (2000). *Physical Causation* (Cambridge: Cambridge University Press).

Dowe, P. (2001). “A Counterfactual Theory of Prevention and ‘Causation’ by Omission”, *Australasian Journal of Philosophy* 79: 216 – 226.

Dowe, P. (2010). “Proportionality and Omissions”, *Analysis* 70: 446 – 451.

Dretske, F. (1972). “Contrastive Statements”, *The Philosophical Review* 81: 411 – 437.

Dretske, F. (1988). *Explaining Behavior* (Cambridge: MIT Press).

Dretske, F. (1996). “Phenomenal Externalism”, *Philosophical Issues, 1: Consciousness* (Atascadero: Ridgeview Publishing).

Dretske, F. and Snyder, A. (1973). “Causality and Sufficiency: Reply to Beauchamp”, *Philosophy of Science* 40: 288 – 291.

Ehring, D. (1996). “Mental Causation, Determinables, and Property Instances”, *Noûs* 4: 461 – 480.

Ehring, D. (1997). *Causation and Persistence* (Oxford: Oxford University Press).

Fair, D. (1979). “Causation and the Flow of Energy”, *Erkenntnis* 14: 219 – 250.

Fodor, J. (1974). “Special Sciences, or the Disunity of Science as a Working Hypothesis”, *Synthese* 28: 97-115.

Fodor, J. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (Cambridge: MIT Press).

Fodor, J. (1989). “Making Mind Matter More”, *Philosophical Topics* 17: 59-79.

Funkhouser, E. (2006). “The Determinate-Determinable Relation”, *Noûs* 40: 548 – 569.

Gibbard, A. (1975). “Contingent Identity”, *Journal of Philosophical Logic* 4: 187 – 221.

Gillet, C. (2002). “The Dimensions of Realization: A Critique of the Standard View”, *Analysis* 64: 316 – 323.

Gillett, C. (2003). “The Metaphysics of Realization, Multiple Realizability, and the Special Sciences”, *Journal of Philosophy* 100: 591 – 603.

Goldman, A. (1969). “The Compatibility of Mechanism and Purpose”, *The Philosophical Review* 78: 468 – 482.

Goldman, A. (1970). *A Theory of Human Action*. Englewood Cliffs, NJ.: Prentice Hall.

Hacker, P.M. (1982). “Events, Ontology and Grammar”, *Philosophy* 57: 477 – 486.

Hall, N. (2004a). “Causation and the Price of Transitivity”, *Causation and Counterfactuals* (Cambridge: MIT Press): 181 – 203.

Hall, N. (2004b). "Two Concepts of Causation", *Causation and Counterfactuals* (Cambridge: MIT Press): 225 – 276.

Hill, C. (1991). *Sensations: A Defense of Type Materialism* (Cambridge: Cambridge University Press).

Hill, C. (1997). "Imaginability, Conceivability, Possibility, and the Mind-Body Problem", *Philosophical Studies* 87: 61 – 85.

Hitchcock, C. (1996a). "Farewell to Binary Causation", *Canadian Journal of Philosophy* 26: 267 – 282.

Hitchcock, C. (1996b). "The Role of Contrast in Causation and Explanation", *Synthese* 107: 395 – 419.

Hitchcock, C. (2001a). "A Tale of Two Effects", *The Philosophical Review* 110: 361 – 396.

Hitchcock, C. (2001b). "The Intransitivity of Causation Revealed in Equations and Graphs", *The Journal of Philosophy* 98: 273 – 299.

Hitchcock, C. (2007). "Prevention, Preemption, and the Principle of Sufficient Reason", *Philosophical Review* 116: 495 – 532.

Horgan, T. (1989). "Mental Quausation", *Philosophical Perspectives* 3: 47 – 76.

Horgan, T. (1993). "From Supervenience to Superdupervenience: Meeting the Demands of a Material World", *Mind* 102: 555-586.

Horgan, T. (1997). "Kim on Mental Causation and Causal Exclusion", *Philosophical Perspectives* 11: 165 – 184.

Horgan, T. (2001). "Causal Compatibilism and the Exclusion Problem", *Theoria* 16: 95 – 116.

Jackson, F. (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. (Oxford: Clarendon Press).

Jackson, F. and Pettit, P. (1990). "Causation in the Philosophy of Mind", *Philosophy and Phenomenological Research* 50: 195 – 214.

Kim, J. (1973). "Causation, Nomic Subsumption and the Concept of Event," *Journal of Philosophy* 70: 217 – 236.

Kim, J. (1976). “Events as Property Exemplifications,” *Action Theory*, M. Brand and D. Walton (eds.) (Dordrecht, Holland: D. Reidel Publishing Co.): 159 – 177.

Kim, J. (1984). “Epiphenomenal and Supervenient Causation”, *Midwest Studies in Philosophy* 9: 257 – 270.

Kim, J. (1989). “Mechanism, Purpose, and Explanatory Exclusion”, *Philosophical Perspectives* 3: 77 – 108.

Kim, J. (1993a). “Multiple Realization and the Metaphysics of Reduction,” *Supervenience and Mind* (Cambridge: Cambridge University Press): 309 – 335.

Kim J. (1993b). “The Myth of Nonreductive Physicalism”, *Supervenience and Mind* (Cambridge: Cambridge University Press): 265 – 284.

Kim, J. (1993c). “Noncausal Connections”, *Supervenience and Mind* (Cambridge: Cambridge University Press): 22 – 32.

Kim, J. (1993e). “Can Supervenience and ‘Non-Strict Laws’ Save Anomalous Monism?”, *Mental Causation*, Heil and Mele (eds.) (Oxford: Clarendon Press): 19 – 26.

Kim, J. (1998). *Mind in a Physical World* (MIT Press).

Kim, J. (2005). *Physicalism, or Something Near Enough* (Princeton University Press).

Kim, J. (2006). “Emergence: Core ideas and issues”, *Synthese*, 151: 547-559.

Kim, J. (2007). “Causation and Mental Causation”, *Contemporary Debates in the Philosophy of Mind*, McLaughlin and Cohen (eds.) (Blackwell Publishing): 228-242.

Kistler, M. (1998). “Reducing Causality to Transmission”, *Erkenntnis* 48: 1 – 24.

Kripke, S. (1971). “Identity and Necessity”, *Identity and Individuation*, Munitz (ed.) (New York University Press): 135 – 164.

Kripke, S. (1980). *Naming and Necessity*. (Cambridge: Harvard University Press).

Levine, J. (2001). *Purple Haze: The Puzzle of Consciousness* (Oxford: Oxford University Press).

Lewis, D. (1966). "An Argument for the Identity Theory", *The Journal of Philosophy* 63: 17 – 25.

Lewis, D. (1971). "Counterparts of Persons and Their Bodies", *Journal of Philosophy* 7: 203 – 211.

Lewis, D. (1973). "Causation", *The Journal of Philosophy* 70: 556 – 567.

Lewis, D. (1979). "Counterfactuals and Time's Arrow", *Noûs* 13: 455 – 476.

Lewis, D. (1980). "Mad Pain, Martian Pain", *Readings in the Philosophy of Psychology*, vol. 1, N. Block (ed.) (Cambridge: Harvard University Press): 216 – 222.

Lewis, D. (1980/2002). "Psychophysical and Theoretical Identifications," *Readings in the Philosophy of Psychology*, vol. 1, N. Block (ed.) (Cambridge: Harvard University Press). Reprinted in *Philosophy of Mind: Classical and Contemporary Readings*, D. Chalmers (ed.): 88 – 94.

Lewis, D. (1983). "New Work for a Theory of Universals", *The Australasian Journal of Philosophy* 61: 343 – 377.

Lewis, D. (1986). *Philosophical Papers*, vol. 1 (Oxford: Oxford University Press).

Lewis, D. (1997). "Finkish Dispositions", *The Philosophical Quarterly* 47: 143 – 158.

Lewis, D. (2004). "Causation as Influence", *Causation and Counterfactuals* (Cambridge: MIT Press): 75 – 106.

List, C. and Menzies, P. (2009). "Non-reductive Physicalism and the Limits of the Exclusion Problem", *Journal of Philosophy* 106: 1 – 21.

Loewer, B. (2001). "From Physics to Physicalism", *Physicalism and its Discontents* (Cambridge: Cambridge University Press): 37 – 56.

Loewer, B. (2002). "Comments on Jaegwon Kim's *Mind in a Physical World*", *Philosophy and Phenomenological Research* 65: 655 – 663.

Loewer, B. (2007). "Mental Causation, or Something Near Enough", *Contemporary Debates in Philosophy of Mind* (Oxford: Blackwell Publishing).

Malcolm, N. (1968). "The Conceivability of Mechanism", *The Philosophical Review* 77: 45 – 72.

Marras, A. (1997). “The Debate on Mental Causation: Davidson and his Critics”, *Dialogue*, 36: 177 – 195.

Marras, A. (1998). “Kim’s Principle of Explanatory Exclusion”, *Australasian Journal of Philosophy* 76: 439 – 451.

McGrath, M. (1998). “Proportionality and Mental Causation: A Fit?”, *Philosophical Perspectives* 12: 167 – 176.

McLaughlin, B. (1989). “Type Epiphenomenalism, Type Dualism, and the Causal Priority of the Physical”, *Philosophical Perspectives* 3: 109 – 135.

McLaughlin, B. (1993). “On Davidson’s Response to the Charge of Epiphenomenalism”, *Mental Causation*, Heil and Mele (eds.) (Oxford: Clarendon Press): 27 – 40.

McLaughlin, B. (2007). “Mental Causation and Shoemaker-Realization”, *Erkenntnis* 67: 149 – 172.

Mellor, D. (1995). *The Facts of Causation* (London: Routledge).

Menzies, P. (1988). “Against Causal Reductionism”, *Mind* 97: 551 – 574.

Menzies, P. (1996). “Probabilistic Causation and the Preemption Problem”, *Mind* 105: 85 – 117.

Menzies, P. (1998). “Are Humean Doubts About Singular Causation Justified?”, *Communication and Cognition* 31: 1 – 26.

Menzies, P. (1999). “Intrinsic versus Extrinsic Conceptions of Causation”, *Laws and Causation: Australasian Studies in the History and Philosophy of Science*, Sankey (ed.) (Dordrecht: Kluwer Academic Publishers): 313 – 329.

Menzies, P. (2003). “The Causal Efficacy of Mental States”, *Physicalism and Mental Causation*, Walter and Heckmann (eds.) (Exeter: Imprint Academic): 195 – 223.

Menzies, P. (2004). “Difference-making in Context”, *Causation and Counterfactuals*, Collins, Hall, and Paul (eds.) (Cambridge: MIT Press): 139 – 180.

Menzies, P. (2007). “Causation in Context”, *Causation, Physics, and the Constitution of Reality: Russell’s Republic Revisited*, H. Price and R. Corry (eds.) (OUP): 191 – 223.

Menzies, P. (2008). "The Exclusion Problem, the Determination Relation, and Contrastive Causation", *Being Reduced: New Essays on Reduction, Explanation, and Causation*, J. Hohwy and J. Kallestrup (eds.) (Oxford: Oxford University Press): 196 – 217.

Merricks, T. (2001). *Objects and Persons* (Oxford: Clarendon Press).

Mills, E. (1996). "Interactionism and Overdetermination", *American Philosophical Quarterly* 33: 105 – 117.

Nagel, E. (1961). *The Structure of Science* (London: Routledge and Keegan Paul).

Olson, E. (1997). *The Human Animal: Personal Identity Without Psychology* (Oxford University Press).

Olson, E. (2002). "Thinking Animals and the Reference of 'I'", *Philosophical Topics* 30: 189 – 208.

Papineau, D. (2001). "The Rise of Physicalism", *Physicalism and its Discontents* (Cambridge: Cambridge University Press): 3 – 36.

Pereboom, D. (2002). "Robust Nonreductive Materialism", *Journal of Philosophy* 99: 499 – 531.

Pereboom, D. (2002). "On Baker's Persons and Bodies", *Philosophy and Phenomenological Research* 64: 615 – 622.

Place, U.T. (1956/2002). "Is Consciousness a Brain Process?", *Philosophy of Mind: Classical and Contemporary Readings*, D. Chalmers (ed.): 55 – 60.

Polger, T. (2004). "Neural Machinery and Realization", *Philosophy of Science* 71: 997 – 1006.

Polger, T. (2008). "Realization and the Metaphysics of Mind", *Australiasian Journal of Philosophy* 85: 233 – 259.

Putnam, H. (1975a). "The Meaning of Meaning", *Philosophical Papers, Vol. II: Mind, Language, and Reality* (Cambridge: Cambridge University Press).

Putnam, H. (1975b). "The Nature of Mental States," *Philosophical Papers, Vol. II: Mind, Language, and Reality* (Cambridge: Cambridge University Press).

Russo, A. and Montminy, M. (forthcoming). "A Defense of Causal Invariantism".

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World* (Princeton: Princeton University Press).

Schaffer, J. (2000a). "Causation by Disconnection", *Philosophy of Science* 67: 285 – 300.

Schaffer, J. (2000b). "Trumping Preemption", *The Journal of Philosophy* 97: 165 – 181.

Schaffer, J. (2001a). "Causation, Influence, and Effluence", *Analysis* 61: 11 – 19.

Schaffer, J. (2001b). "Causes as Probability Raisers of Processes", *The Journal of Philosophy* 98: 75 – 92.

Schaffer, J. (2003). "Overdetermining Causes", *Philosophical Studies* 114: 23 – 45.

Schaffer, J. (2004a). "Causes Need Not be Physically Connected to their Effects: The Case for Negative Causation", *Contemporary Debates in the Philosophy of Science* (Basil Blackwell): 197 – 216.

Schaffer, J. (2004b). "Of Ghostly and Mechanical Events", *Philosophy and Phenomenological Research* 68: 230 – 244.

Schaffer, J. (2005). "Contrastive Causation", *The Philosophical Review* 114: 327 – 358.

Schaffer, J. (2010). "Contrastive Causation in Law", *Logical Theory* 16: 259 – 297.

Schaffer, J. (forthcoming). "Causal Contextualism".

Schiffer, S. (1987). *Remnants of Meaning* (Cambridge: MIT Press).

Shapiro, L. (2000). "Multiple Realizations", *The Journal of Philosophy* 97: 635 – 654.

Shapiro, L. (2004). *The Mind Incarnate* (Cambridge: MIT Press).

Shoemaker, S. (2001). "Realization and Mental Causation", *Physicalism and its Discontents* (Cambridge: Cambridge University Press): 74 – 98.

Shoemaker, S. (2003a). "Causality and Properties", *Identity, Cause, and Mind* (Oxford: Clarendon Press): 206 – 233.

Shoemaker, S. (2003b). “Identity, Properties, and Causality”, *Identity, Cause, and Mind* (Oxford: Clarendon Press): 234 – 260.

Shoemaker, S. (2003c). “Realization, Micro-Realization, and Coincidence”, *Philosophy and Phenomenological Research* 67: 1 – 23.

Shoemaker, S. (2007). *Physical Realization* (Oxford: Oxford University Press).

Sider, T. (2002). “Review of Lynne Rudder Baker, *Persons and Bodies*”, *Journal of Philosophy* 99: 45 – 48.

Sider, T. (2003). “What’s So Bad About Overdetermination?”, *Philosophy and Phenomenological Research* 67: 719 – 726.

Simona, A. (2011). “Counterfactuals, Overdetermination, and Mental Causation”, *Proceedings from the Aristotelian Society* 3: 469 – 477.

Sober, E. (1999). “The Multiple Realizability Argument Against Reductionism,” *Philosophy of Science* LXVI: 542 – 564.

Sosa, E. (1984). “Mind-Body Interaction and Supervenient Causation”, *Midwest Studies in Philosophy* 9: 271 – 281.

Sosa, E. (1993). “Davidson’s Thinking Causes”, *Mental Causation*, Heil and Mele (eds.) (Oxford: Clarendon Press): 41 – 50.

Smart, J.C.C. (1959/2002). “Sensations and Brain Processes”, *Philosophy of Mind: Classical and Contemporary Readings*, D. Chalmers (ed.): 60 – 68.

Stoljar, D. (2008). “Distinctions in Distinction”, *Being Reduced: New Essays on Reduction, Explanation, and Causation*, Jakob Hohwy and Jesper Kallestrup (eds.) (Oxford: Oxford University Press): 263 – 279.

Sturgeon, S. (1998). “Physicalism and Overdetermination”, *Mind* 107: 411 – 432.

Thomasson, A. (1998). “A Nonreductivist Solution to Mental Causation”, *Philosophical Studies* 89: 181 – 195.

Thomson, J. (1997). Parts: A Study in Ontology (Oxford: Clarendon Press).

Thomson, J. (1998). “The Statue and the Clay”, *Noûs* 32: 149 – 173.

Tye, M. (1995). *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind* (Cambridge: MIT Press).

Tye, M. (2000). *Consciousness, Color, and Content* (Cambridge: MIT Press).

Van Inwagen, P. (1990). *Material Beings* (Ithaca: Cornell University Press).

Wasserman, R. (2002). "The Standard Objection to the Standard Account", *Philosophical Studies* 111: 197 – 216.

Wasserman, R. (2004). "The Constitution Question", *Noûs* 38: 693 – 710.

Wiggens, D. (1968). "On Being in the Same Place at the Same Time", *Philosophical Review* 77: 90 – 95.

Wiggens, D. (1980). *Sameness and Substance* (Cambridge: Harvard University Press).

Woodward, J. (1984). "A Theory of Singular Causal Explanation", *Erkenntnis* 37: 231 – 262.

Woodward, J. (2003). *Making Things Happen* (Oxford: Oxford University Press).

Woodward, J. (2008). "Mental Causation and Neural Mechanisms", *Being Reduced: New Essays on Reduction, Explanation, and Causation*, Hohwy and Kallestrup (eds.) (Oxford: Oxford University Press): 218 – 262.

Yablo, S. (1987). "Identity, Essence, and Indiscernibility", *The Journal of Philosophy* 84: 293 – 314.

Yablo, S. (1992a). "Cause and Essence", *Synthese* 3: 403 – 449.

Yablo, S. (1992b). "Mental Causation", *The Philosophical Review* 101: 245-280.

Yablo, S. (1997). "Wide Causation", *Philosophical Perspectives* 11: 251 – 281.

Yablo, S. (2007). "The Seven Habits of Highly Effective Thinkers". Retrieved October 26th, 2011 from Stephen Yablo's MIT website: http://mit.edu/~yablo/effthink.html

Zimmerman, D. (1998). "Criteria of Identity and the 'Identity Mystics'", *Erkenntnis* 48: 281 – 301.

Zimmerman, D. (2002). "Persons and Bodies: Constitution without Mereology?", *Philosophy and Phenomenological Research* 64: 599 – 606.